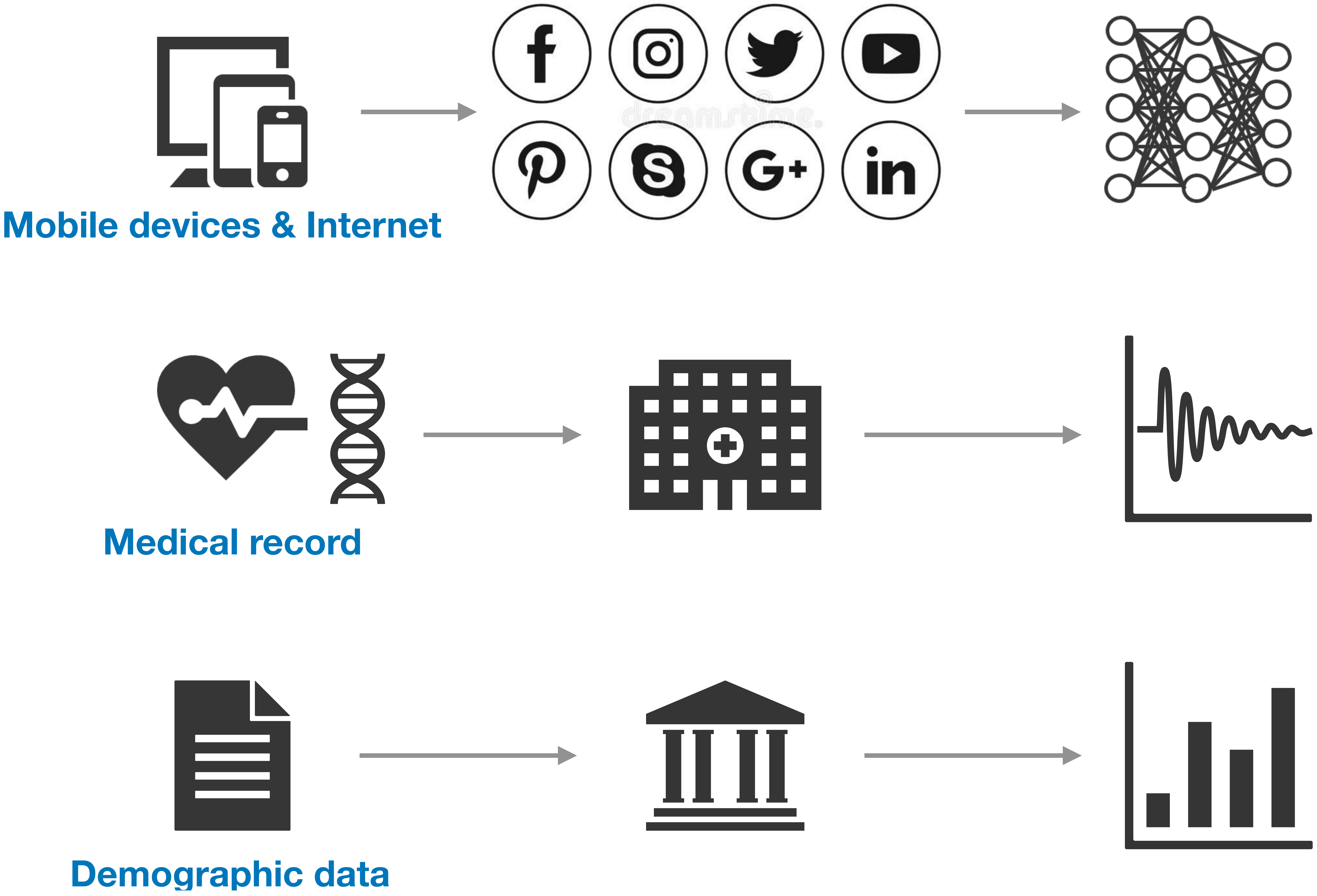
# Privacy-preserving Generative Modeling

**Presenter:** Dingfan Chen
**Supervisor:** Prof. Dr. Mario Fritz
**Affiliation:** CISPA – Helmholtz Center for Information Security

# Privacy-preserving Generative Modeling

# Data Privacy in ML:

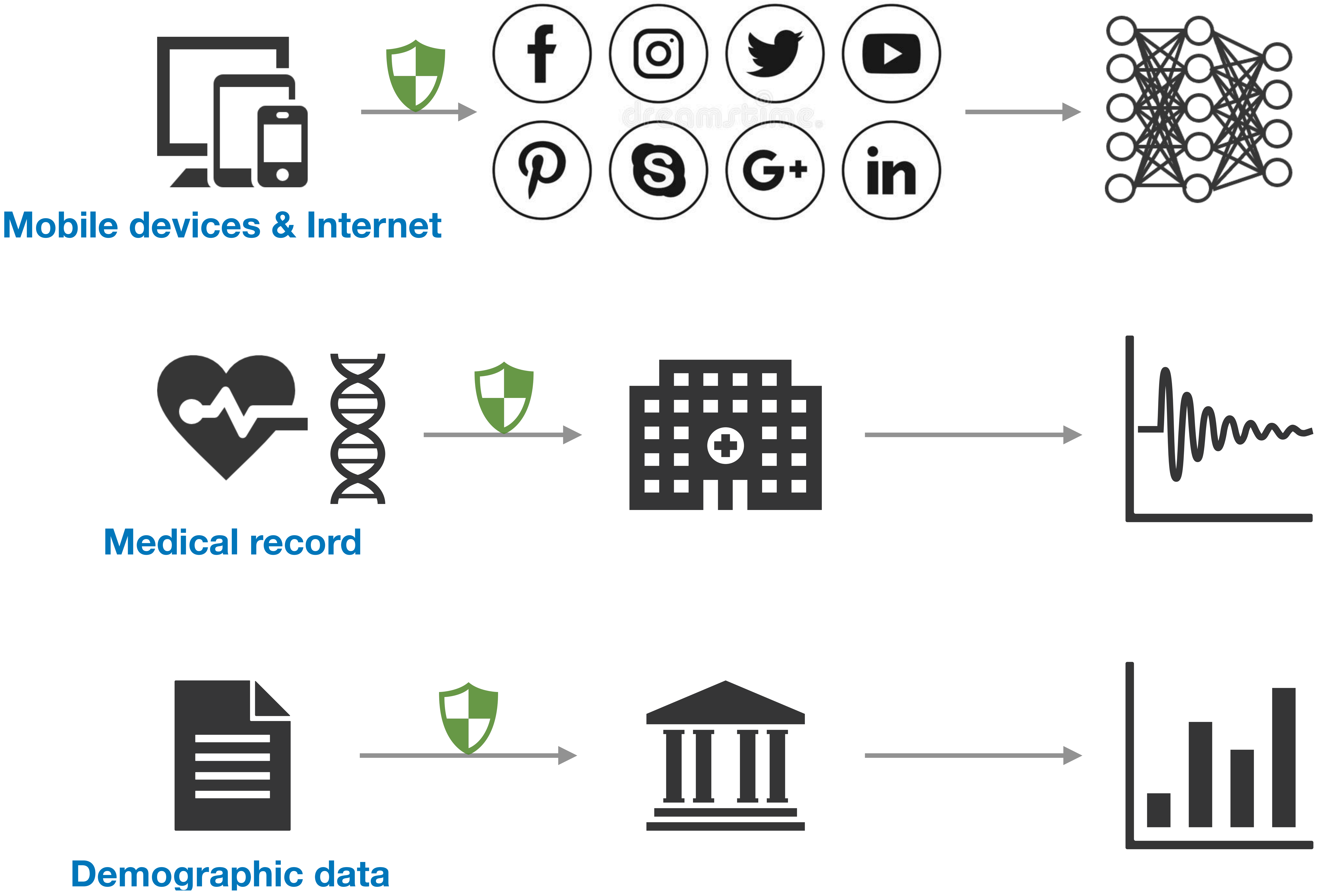
- Sensitive data is **ubiquitous**



# Data Privacy in ML:

- Sensitive data is **ubiquitous**

Our task: **Data sanitization**



# Data Privacy in ML:

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

- Protecting privacy is **non-trivial**
  - **Anonymization** vs. **Deidentification**

    (ZIP code, date of birth, gender) is sufficient to identify 87% of US population[1,2]

**Anonymous medical data**

| ID | QID | | | SA |
|---|---|---|---|---|
| Name | ZIP code | Age | Sex | Disease |
| Alice | 47677 | 29 | F | Ovarian Cancer |
| Betty | 47602 | 22 | F | Ovarian Cancer |
| Charles | 47678 | 27 | M | Prostate Cancer |
| David | 47905 | 43 | M | Flu |
| Emily | 47909 | 52 | F | Heart Disease |
| Fred | 47906 | 47 | M | Heart Disease |

**Voter registration data**

| Name | ZIP code | Age | Sex |
|---|---|---|---|
| Alice | 47677 | 29 | F |
| Bob | 47983 | 65 | M |
| Carol | 47677 | 22 | F |
| Dan | 47532 | 23 | M |
| Ellen | 46789 | 43 | F |
| Fabian | 47905 | 30 | M |

[1] Golle, Philippe. "Revisiting the uniqueness of simple demographics in the US population.", *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society,* 2006.
[2] Sweeney, L., "K-anonymity: A model for protecting privacy.", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002

# Data Privacy in ML:

- Protecting privacy is **non-trivial**
  - **Anonymization** vs. **Deidentification**

    (ZIP code, date of birth, gender) is sufficient to identify 87% of US population[1,2]

**Anonymous medical data**

| QID | | | SA |
|---|---|---|---|
| ZIP code | Age | Sex | Disease |
| 47677 | 29 | F | Ovarian Cancer |
| 47602 | 22 | F | Ovarian Cancer |
| 47678 | 27 | M | Prostate Cancer |
| 47905 | 43 | M | Flu |
| 47909 | 52 | F | Heart Disease |
| 47906 | 47 | M | Heart Disease |

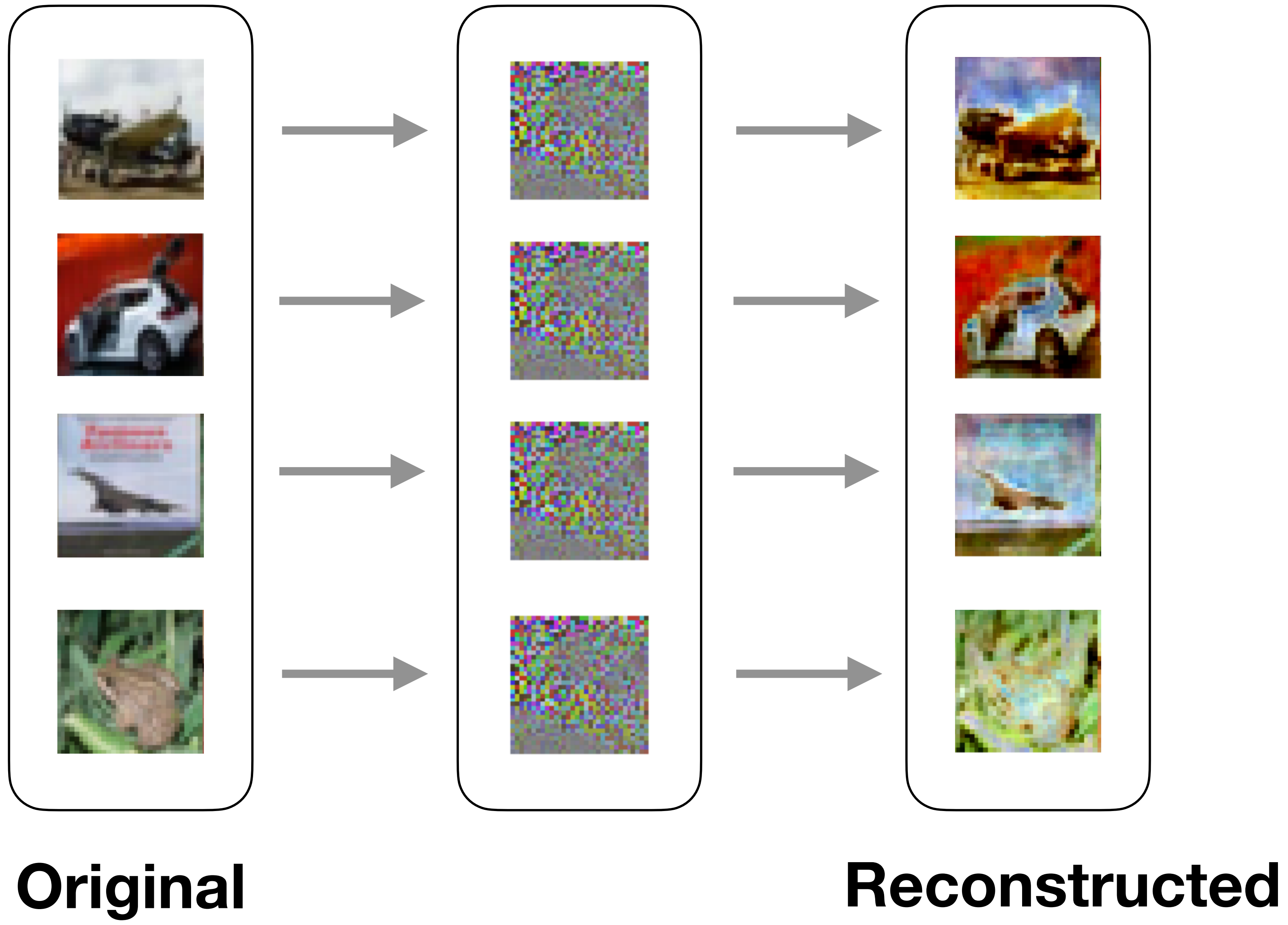
**Voter registration data**

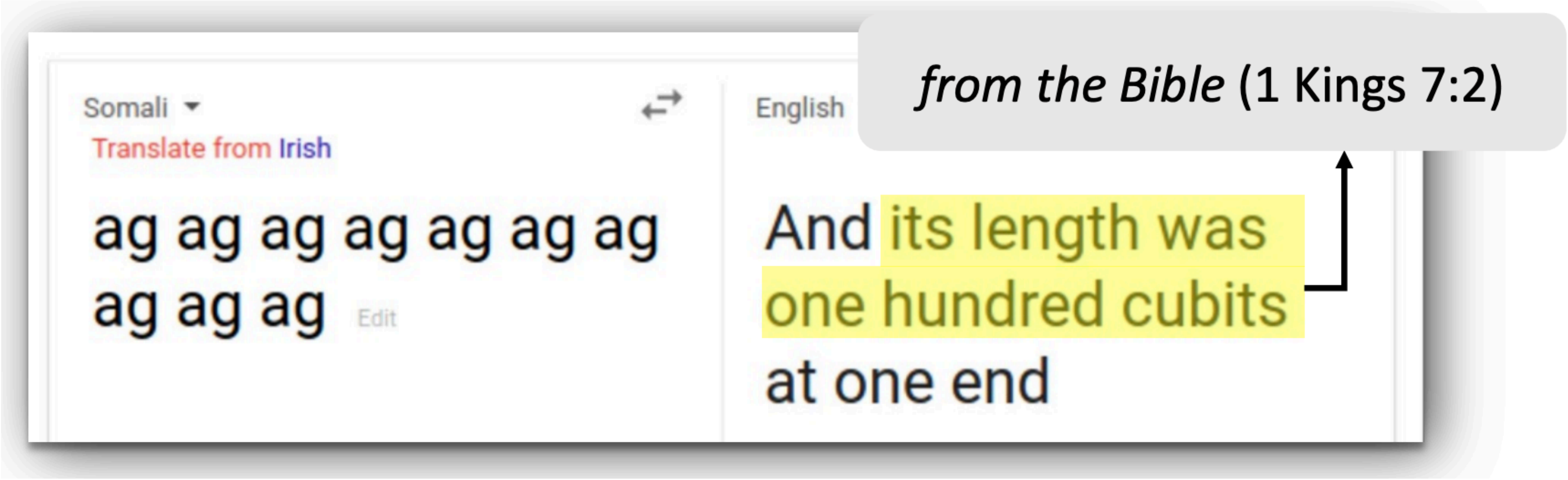
| Name | ZIP code | Age | Sex |
|---|---|---|---|
| Alice | 47677 | 29 | F |
| Bob | 47983 | 65 | M |
| Carol | 47677 | 22 | F |
| Dan | 47532 | 23 | M |
| Ellen | 46789 | 43 | F |
| Fabian | 47905 | 30 | M |

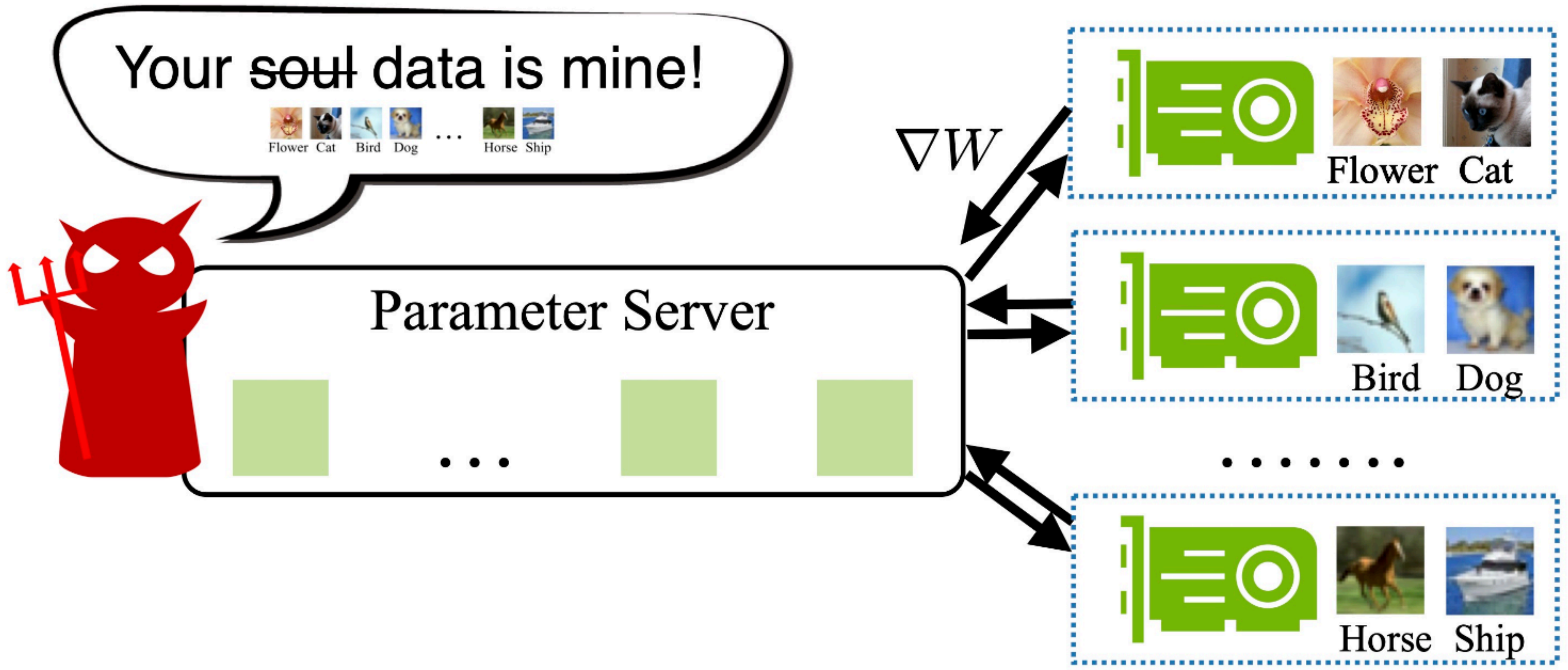
[1] Golle, Philippe. "Revisiting the uniqueness of simple demographics in the US population.", *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society,* 2006.
[2] Sweeney, L., "K-anonymity: A model for protecting privacy.", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002

# **Data Privacy in ML:**

- Protecting privacy is **non-trivial**
  - **Reconstruction** from features, models, gradients, etc.




[1] Carlini, Nicholas, et al. "Is Private Learning Possible with Instance Encoding?." *IEEE Security & Privacy, 2021*.
[2] Carlini, Nicholas, et al. "Extracting training data from large language models." *USENIX Security 21*, 2021.
[3] Zhu, Ligeng,et al. "Deep leakage from gradients.", *NeurIPS*, 2019.

# Rigorous Privacy Guarantee

[1] Dwork. “Differential privacy.”, Automata, languages and programming, 2006
[2] Mironov, Ilya, “Renyi Differential Privacy”, *CSF,* 2017

# Rigorous Privacy Guarantee

## Differential privacy (DP)[1]

- Belonging to a dataset ≈ Not belonging to it
- A mechanism $\mathscr{A}$ is $(\varepsilon, \delta)$-DP iff for any **neighboring datasets** $D$ and $D'$ differing in a single data point, and any $S \subseteq \mathrm{range}(\mathscr{A})$, we have:

$$\Pr[\mathscr{A}(D) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathscr{A}(D') \in S] + \delta$$

- **Bound the maximal influence** of each individual, introduce **randomness**
- Currently, people always turn it into bounding the divergence[2]:

$$D_{\alpha}(\mathscr{A}(D) \| \mathscr{A}(D')) \leq \varepsilon$$

**Properties**

- Allows **quantifying** the privacy risk
- Compose gracefully for iterative methods
- Closed under **post-processing**

[1] Dwork. "Differential privacy.", Automata, languages and programming, 2006
[2] Mironov, Ilya, "Renyi Differential Privacy", *CSF,* 2017

# Rigorous Privacy Guarantee

## Differential privacy (DP)[1]

- Belonging to a dataset ≈ Not belonging to it
- A mechanism $\mathscr{A}$ is $(\varepsilon, \delta)$-DP iff for any **neighboring datasets** $D$ and $D'$ differing in a single data point, and any $S \subseteq \mathrm{range}(\mathscr{A})$, we have:

$$\Pr[\mathscr{A}(D) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathscr{A}(D') \in S] + \delta$$

- **Bound the maximal influence** of each individual, introduce **randomness**
- Currently, people always turn it into bounding the divergence[2]:

$$D_{\alpha}(\mathscr{A}(D) \| \mathscr{A}(D')) \leq \varepsilon$$

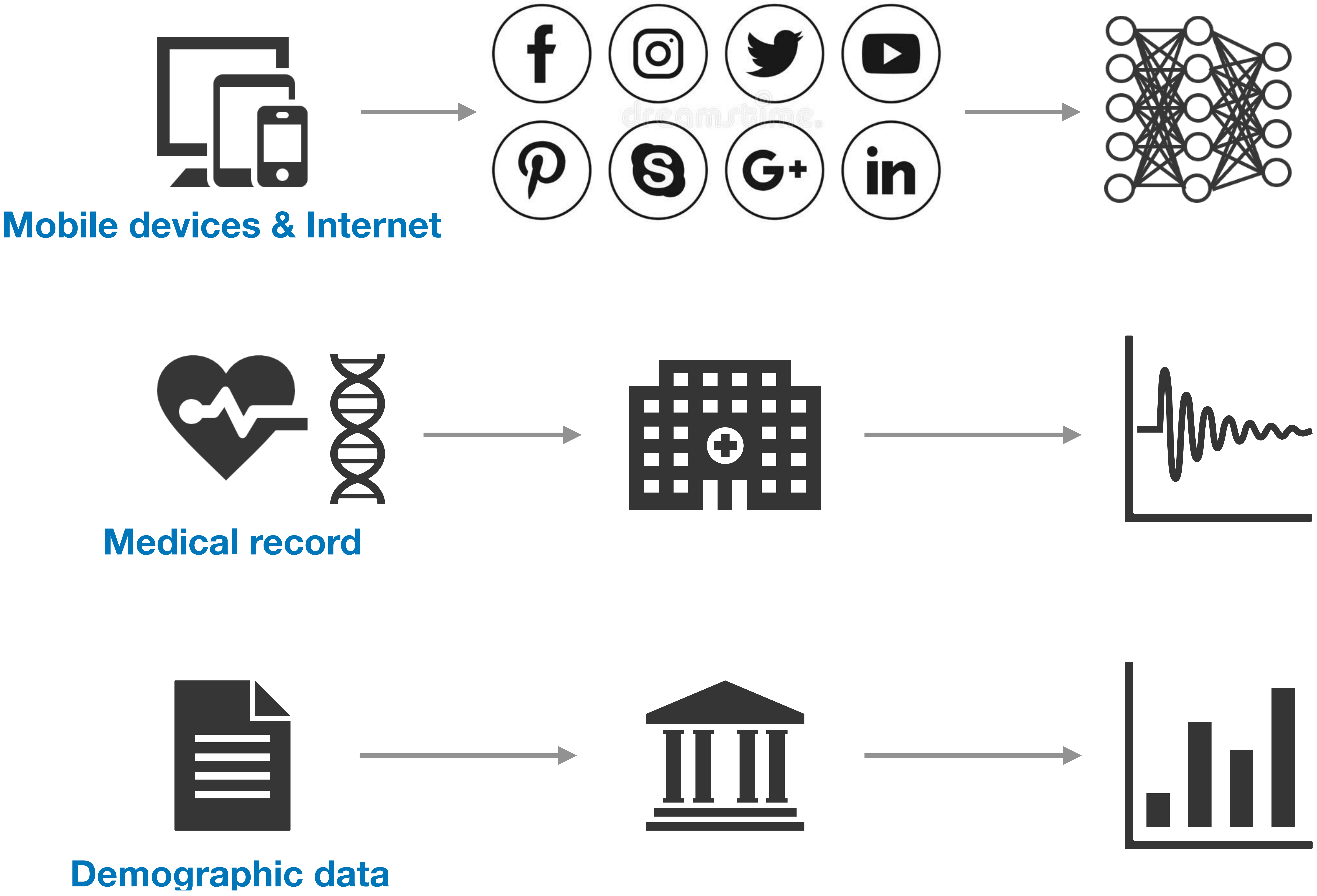
**Properties**

- Allows **quantifying** the privacy risk
- Compose gracefully for iterative methods
- Closed under **post-processing**

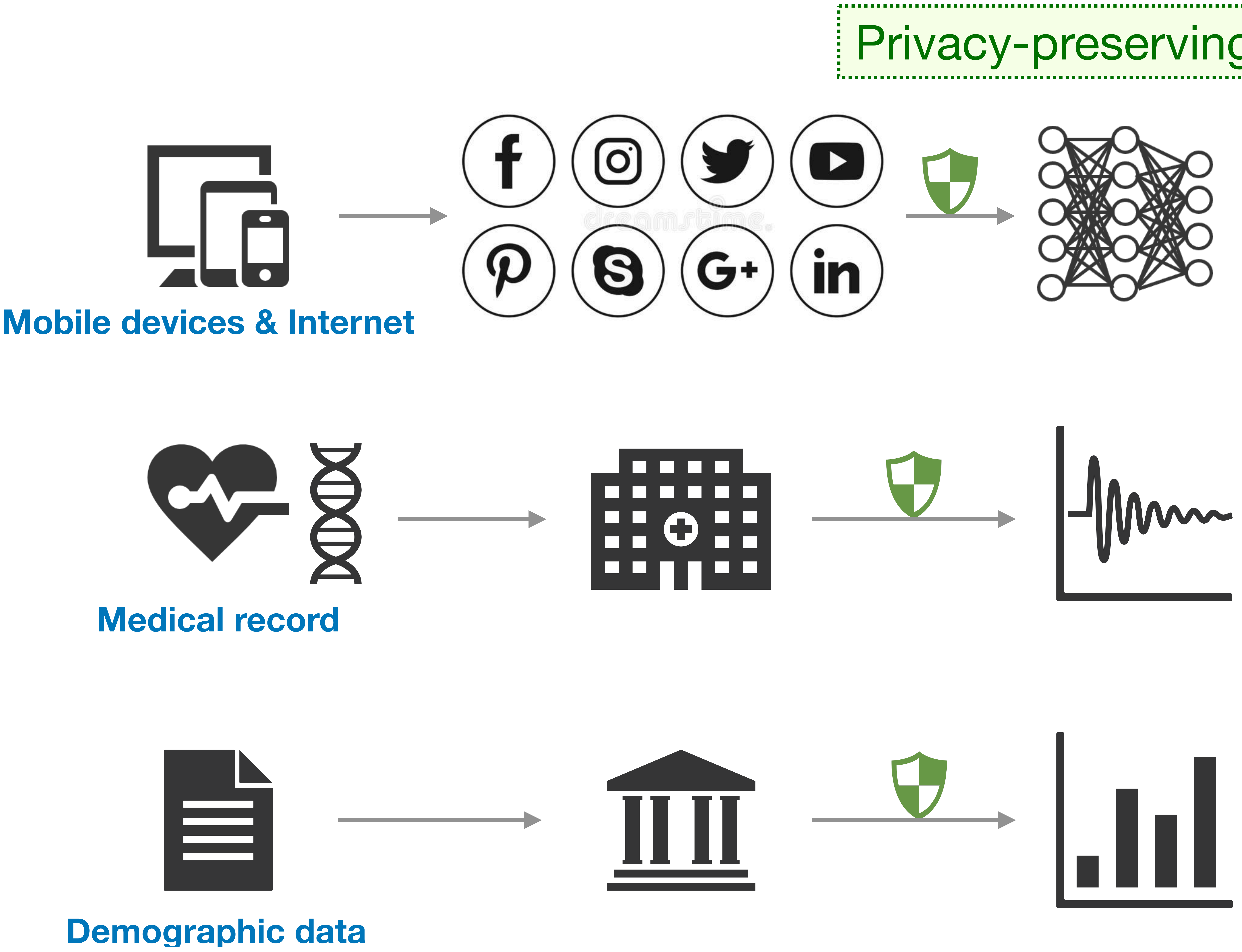
Sanitized data can be freely used for downstream analysis

[1] Dwork. "Differential privacy.", Automata, languages and programming, 2006
[2] Mironov, Ilya, "Renyi Differential Privacy", *CSF,* 2017

# Privacy-preserving Generation vs. Analysis

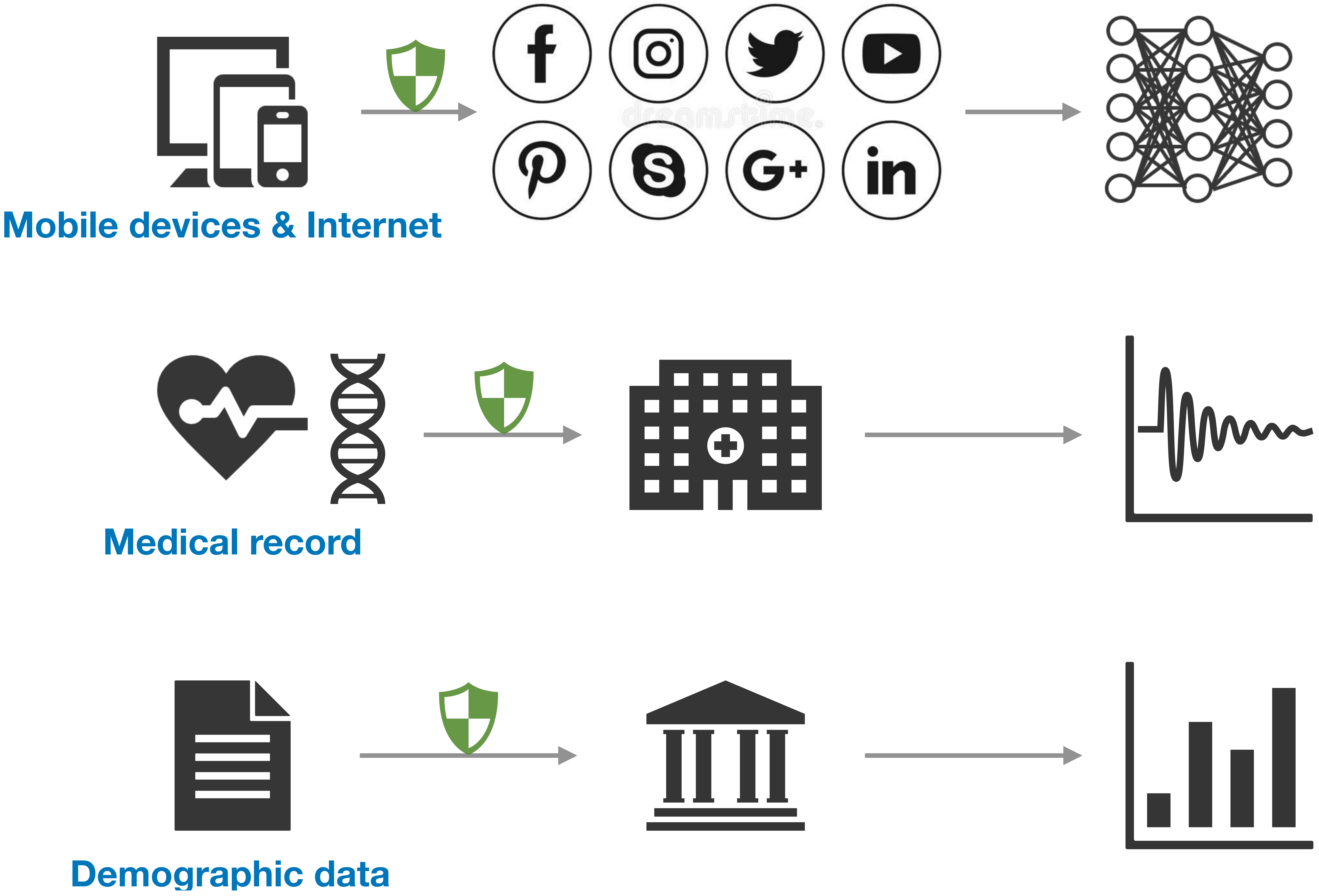


# Privacy-preserving Generation vs. Analysis



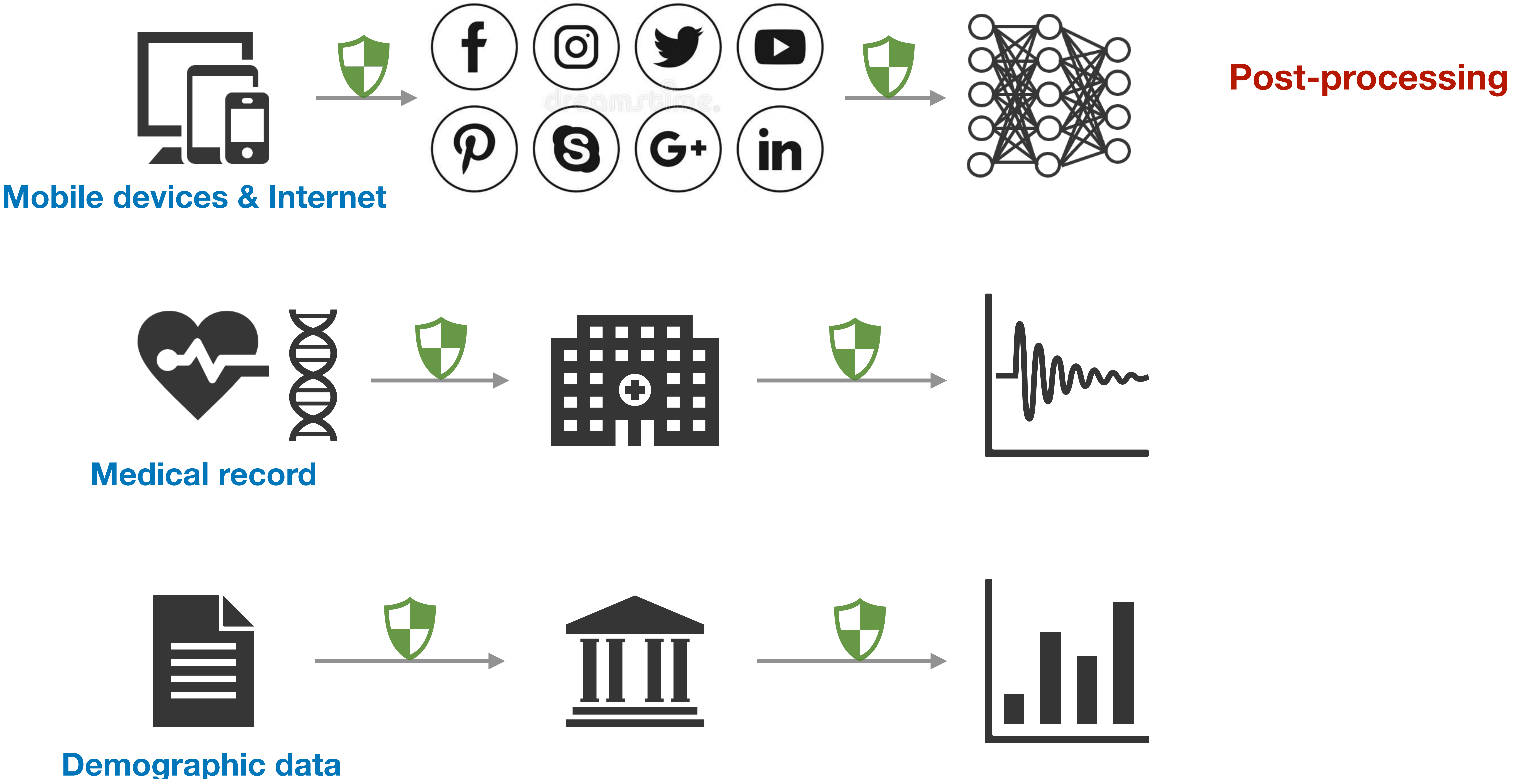
# Privacy-preserving Generation vs. Analysis

Our task: **Data sanitization**



# Privacy-preserving Generation vs. Analysis

Our task: **Data sanitization**



# How to Train a Model under DP?

**Algorithm 1** Differentially private SGD (Outline)

**Input:** Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate $\eta_t$, noise scale $\sigma$, group size $L$, gradient norm bound $C$.

**Initialize** $\theta_0$ randomly

**for** $t \in [T]$ **do**

Take a random sample $L_t$ with sampling probability $L/N$

**Compute gradient**

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

**Clip gradient**

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

**Add noise**

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

**Descent**

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

**Output** $\theta_T$ and compute the overall privacy cost $(\varepsilon, \delta)$ using a privacy accounting method.

# How to Train a Model under DP?

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

**Algorithm 1** Differentially private SGD (Outline)

**Input:** Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate $\eta_t$, noise scale $\sigma$, group size $L$, gradient norm bound $C$.

**Initialize** $\theta_0$ randomly

**for** $t \in [T]$ **do**

Take a random sample $L_t$ with sampling probability $L/N$

**Compute gradient**

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

**Clip gradient**

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

**Add noise**

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

**Descent**

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

**Output** $\theta_T$ and compute the overall privacy cost $(\varepsilon, \delta)$ using a privacy accounting method.

$$\mathcal{M}(D) \triangleq f(D) + \mathcal{N}(0, S_f^2 \cdot \sigma^2)$$

$\mathcal{M}$ : Gaussian Mechanism

$D$ : Dataset

$f$ : Real-valued function

$S_f$ : Sensitivity

$\sigma$ : Noise scale

$$S_f = \max_{D,D'} \|f(D) - f(D')\|_2$$

**Sensitivity**

$$S_f = \max_{x_i} \|\bar{\mathbf{g}}_t(x_i)\|_2 = C$$

# Privacy-preserving Generative Modeling

# Generative Models

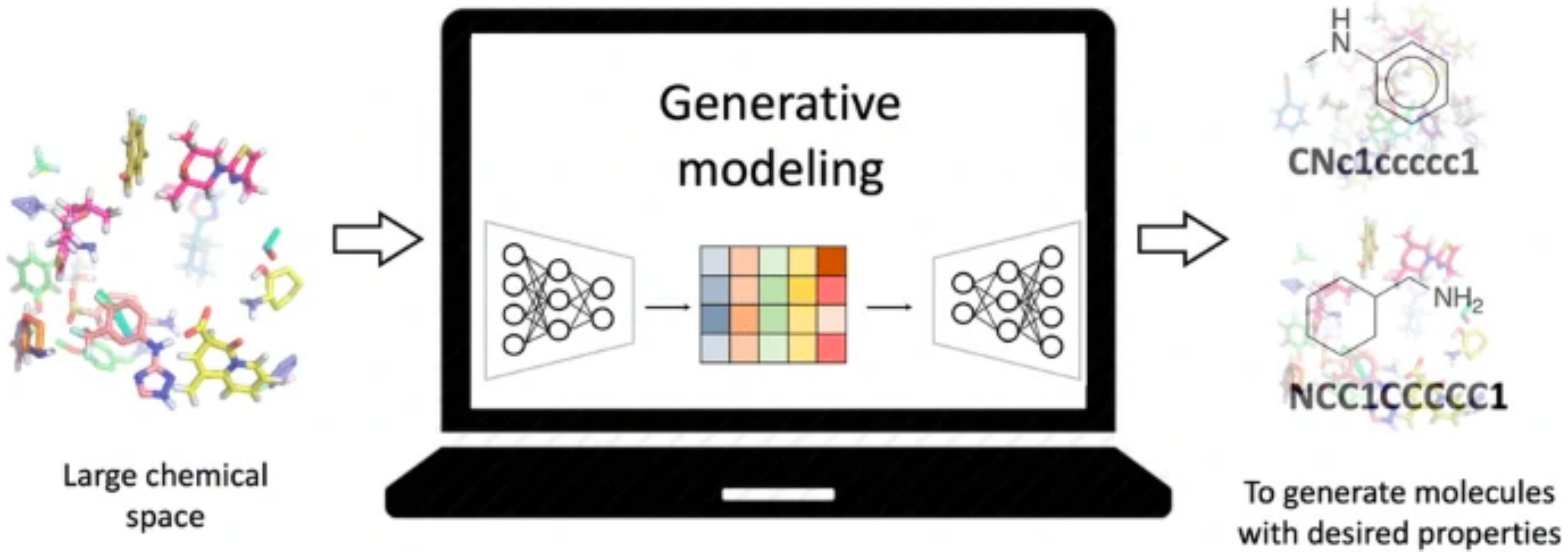




Image source: “Generative chemistry: drug discovery with deep learning generative models”

“An armchair in the shape of an avocado”

“A snail made of a harp”

Image source: https://openai.com/blog/dall-e/

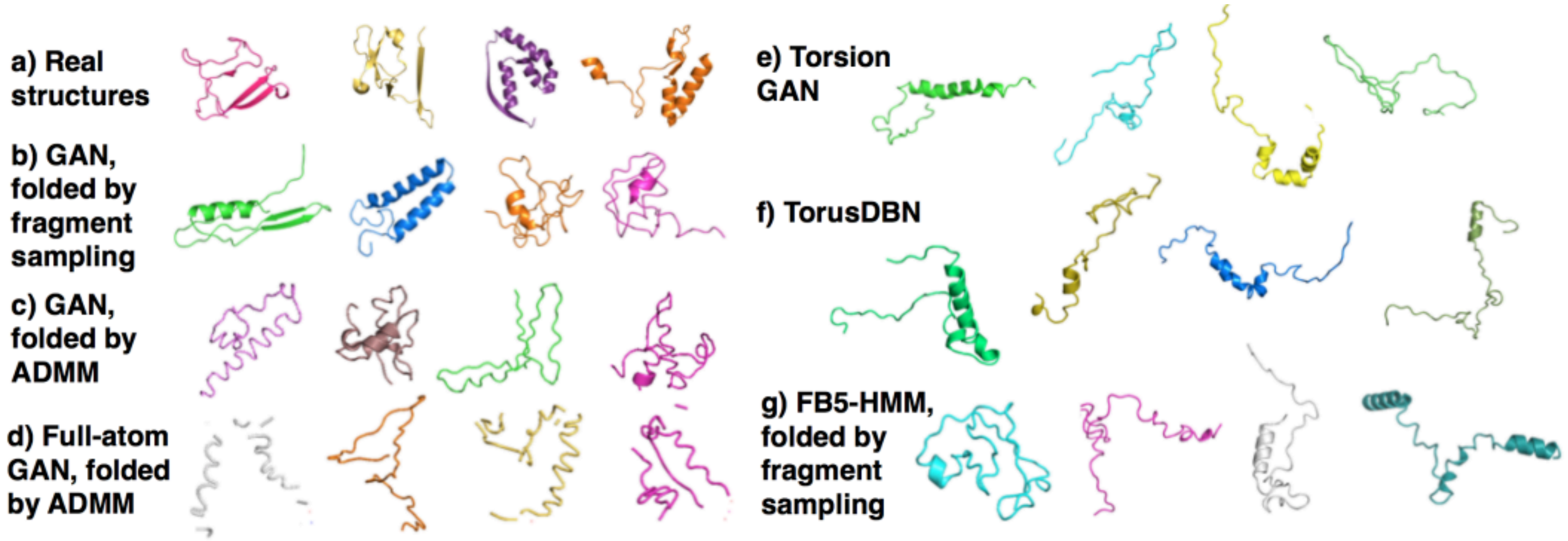



Image source: “Generative Modeling for Protein Structures”

# Generative Models

- **Overview:**
  - **GAN**: Adversarial training
  - **VAE**: Maximize variational lower bound
  - **Flow-based**: Invertible transforms of distributions
  - **Diffusion models**: Gradually add Gaussian noise and reverse

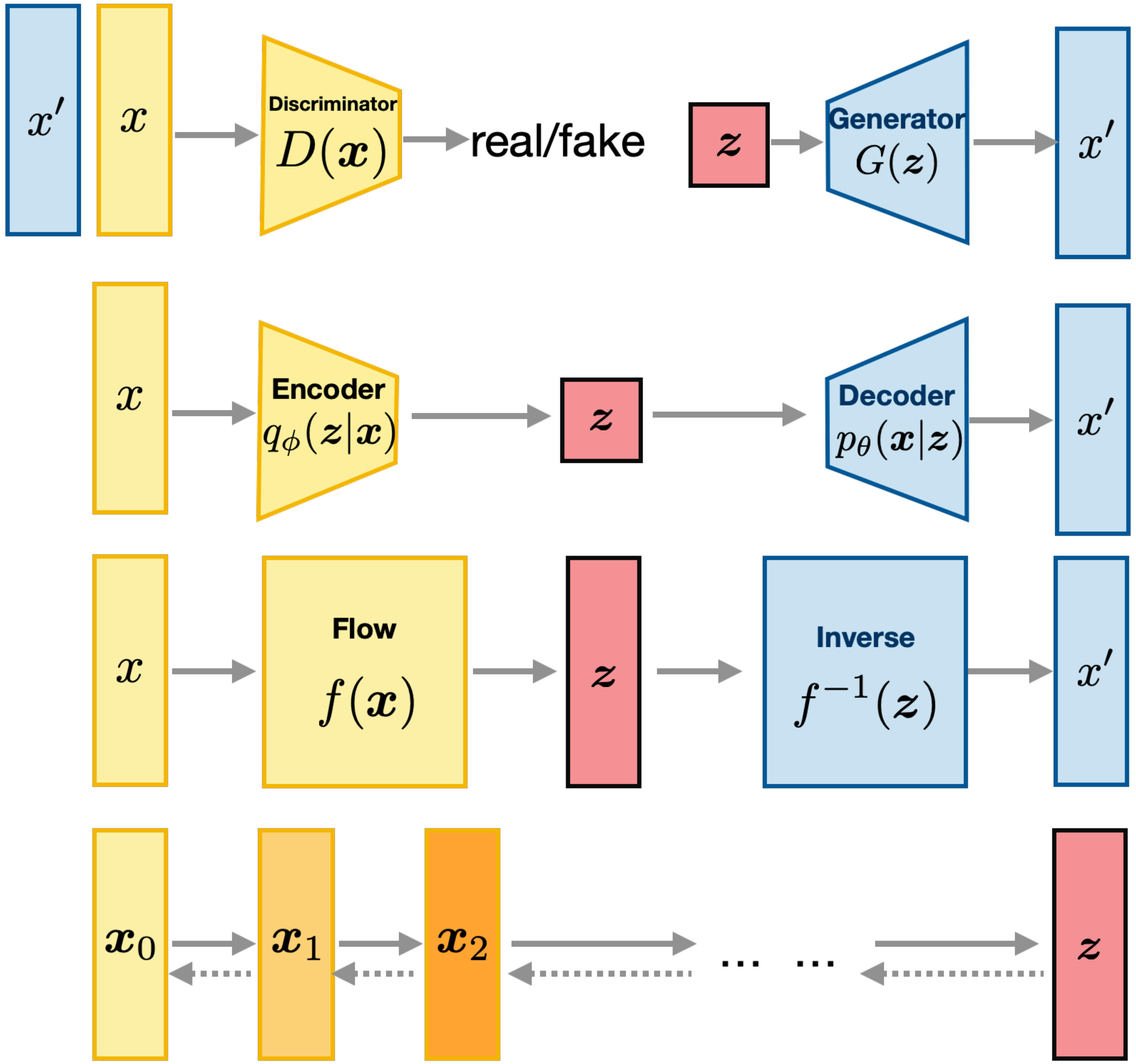
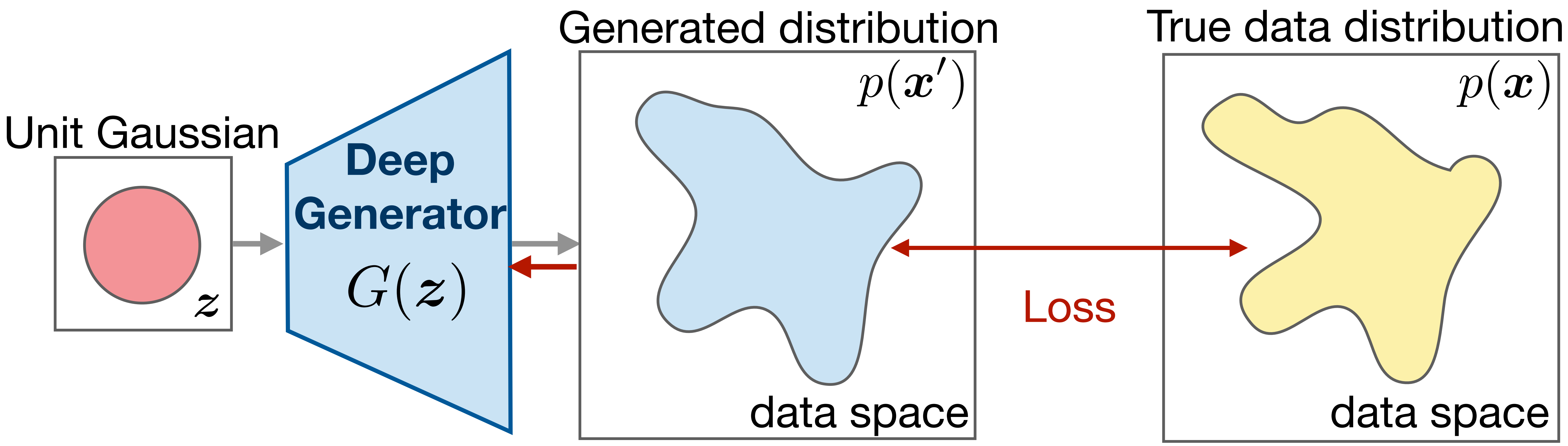



Image source: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

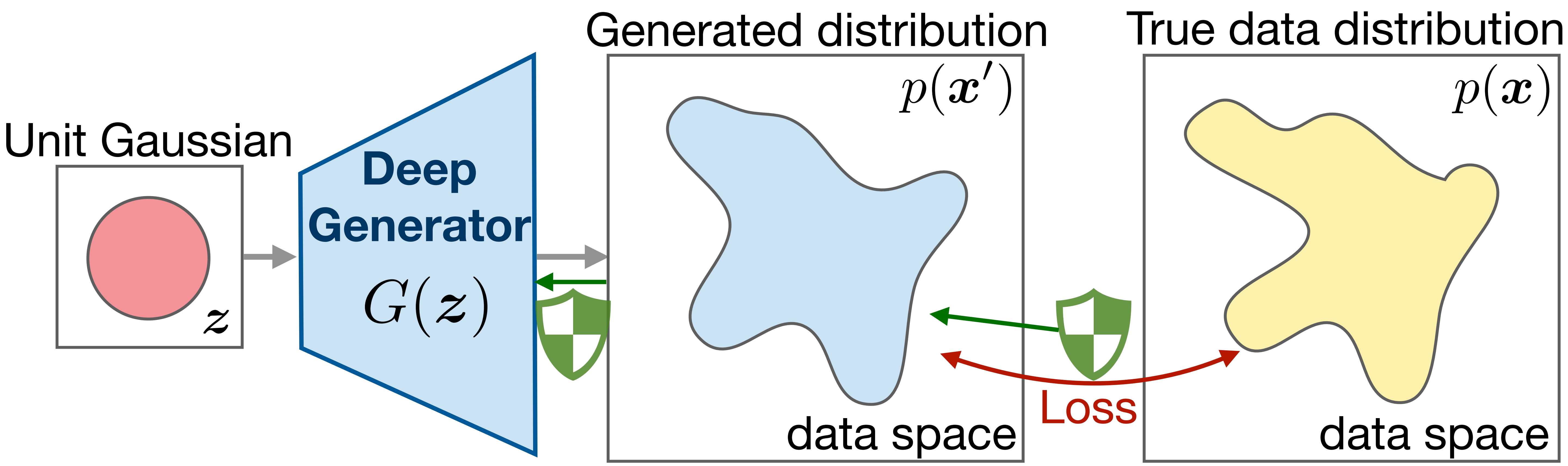
# Generative Models

- **Overview:**
  - Latent variable model $\boldsymbol{z} \rightarrow \boldsymbol{x}$
  - Learn a mapping from simple distribution $p(\boldsymbol{z})$ to complex data distribution



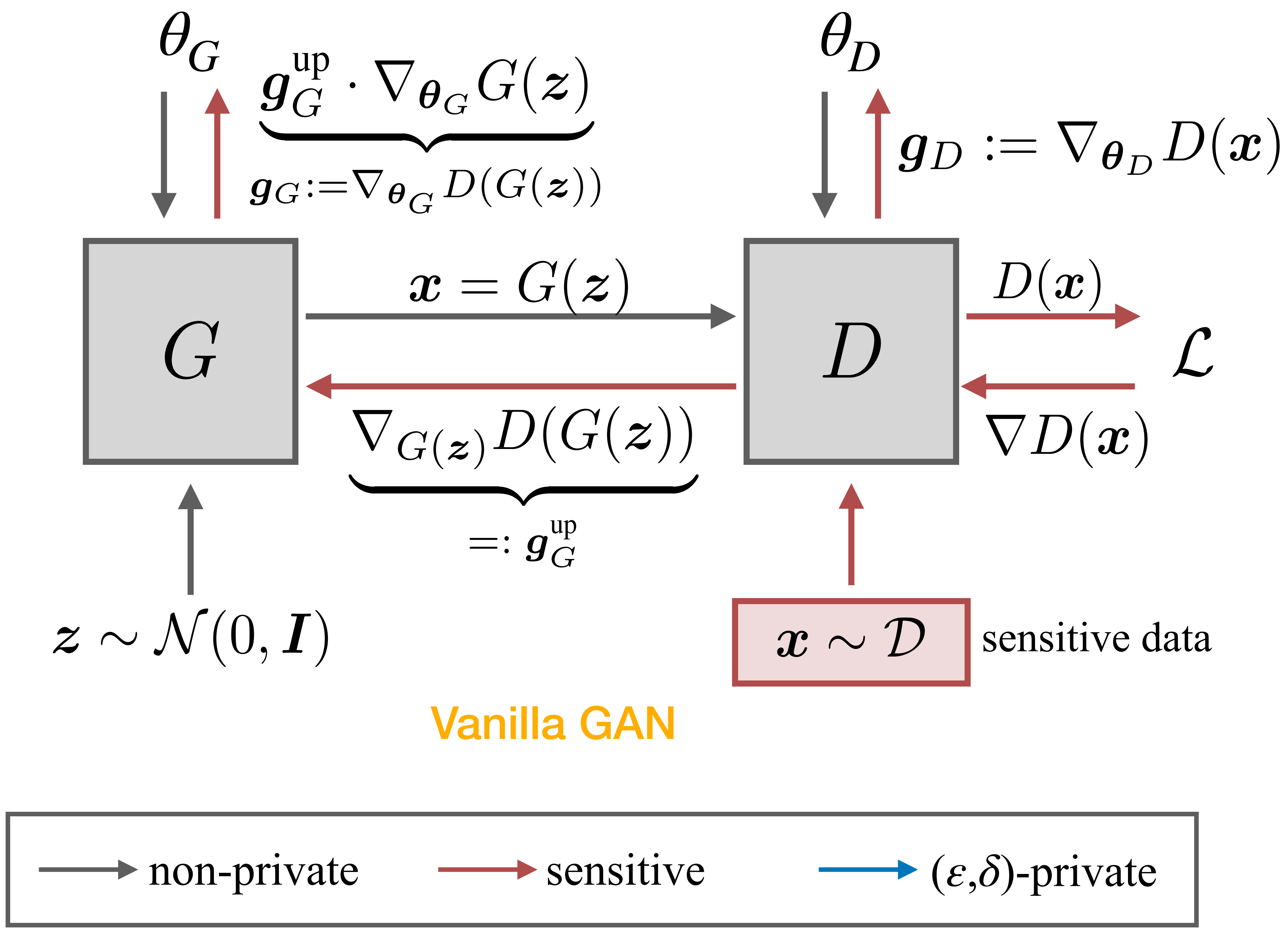
# Privacy-preserving Data Generation

- **Task:**
  - Learn to generate high-dimensional **sanitized data**
- **Key:**
  - Rigorous privacy guarantee → Differential Privacy
  - High-dimensional data → Deep Neural Networks
  - General purpose → Generality & Expressiveness
- **Overview:**



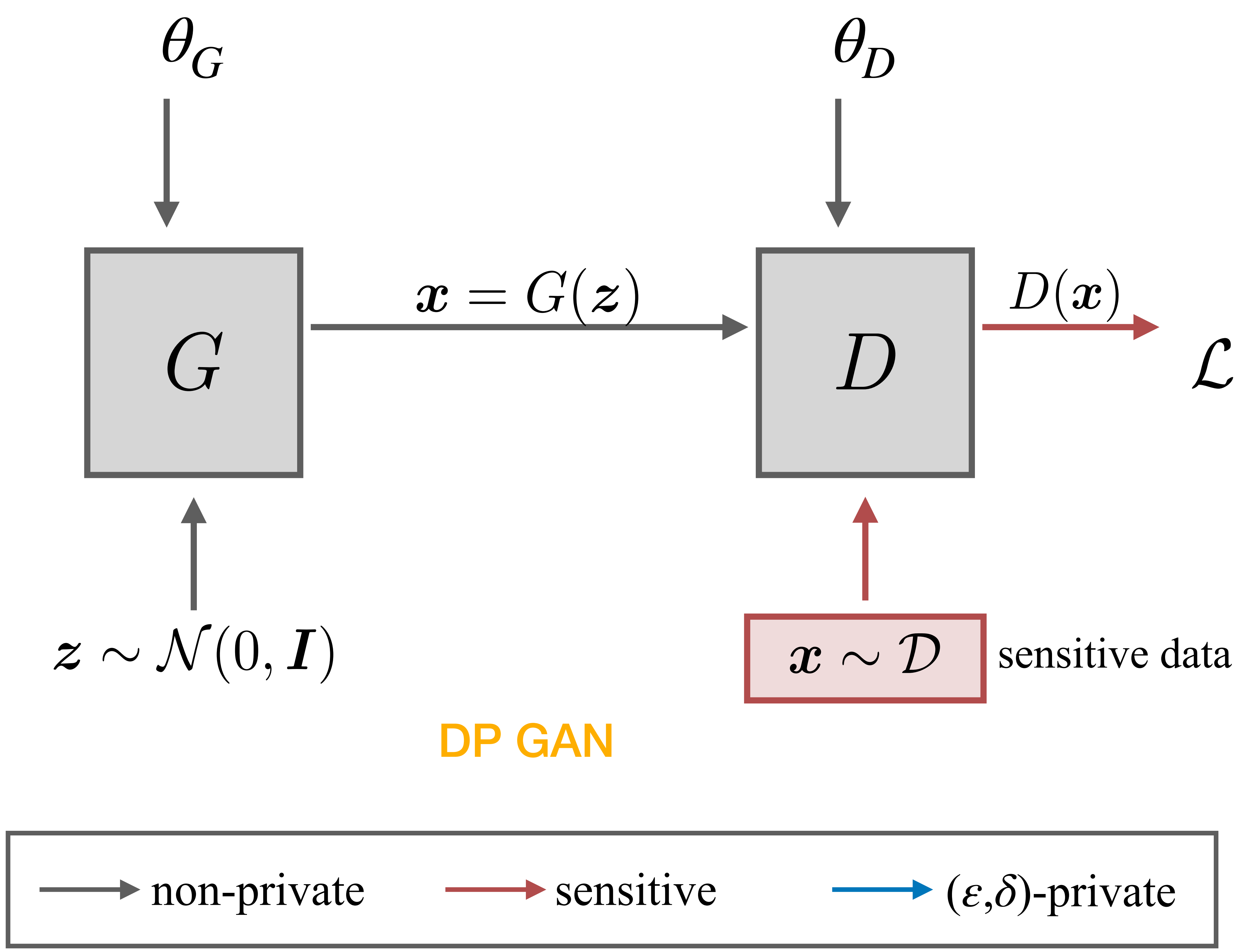
# Existing Solutions

- **Generative adversarial networks (GANs):**
  - Gradient
    $$\boldsymbol{g}^{(t)} := \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_D, \boldsymbol{\theta}_G)$$
  - Gradient descent step
    $$\boldsymbol{\theta}^{(t+1)} := \boldsymbol{\theta}^{(t)} - \eta \cdot \boldsymbol{g}^{(t)}$$



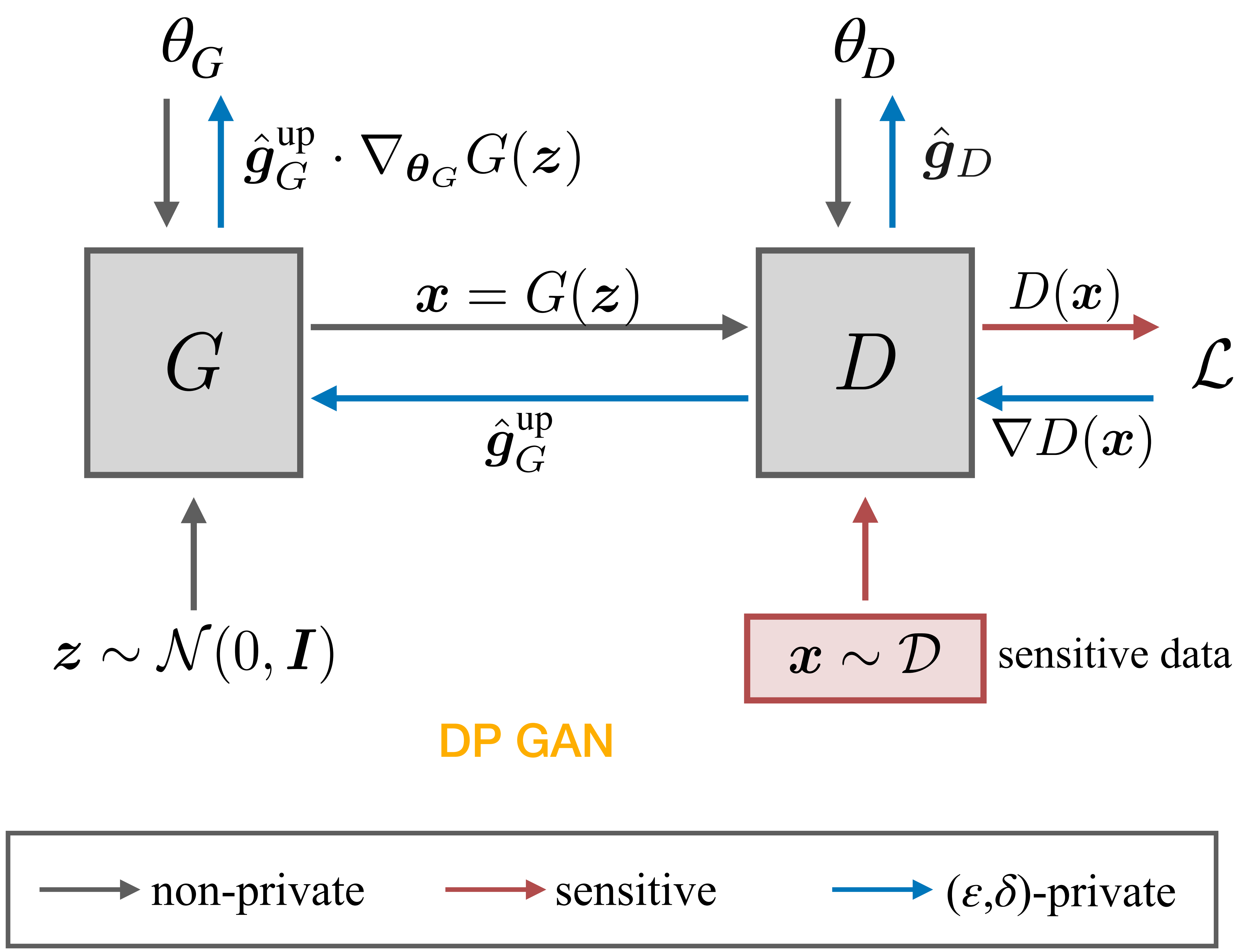
# Existing Solutions

- **Generative adversarial networks (GANs):**
  - Gradient
    $$\boldsymbol{g}^{(t)} := \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_D, \boldsymbol{\theta}_G)$$
  - Sanitization mechanism
    $$\hat{\boldsymbol{g}}^{(t)} := \mathcal{M}_{\sigma,C}(\boldsymbol{g}^{(t)}) = \mathrm{clip}(\boldsymbol{g}^{(t)}, C) + \mathcal{N}(0, \sigma^2 C^2 \boldsymbol{I})$$
  - Gradient descent step
    $$\boldsymbol{\theta}^{(t+1)} := \boldsymbol{\theta}^{(t)} - \eta \cdot \hat{\boldsymbol{g}}^{(t)}$$



# Existing Solutions

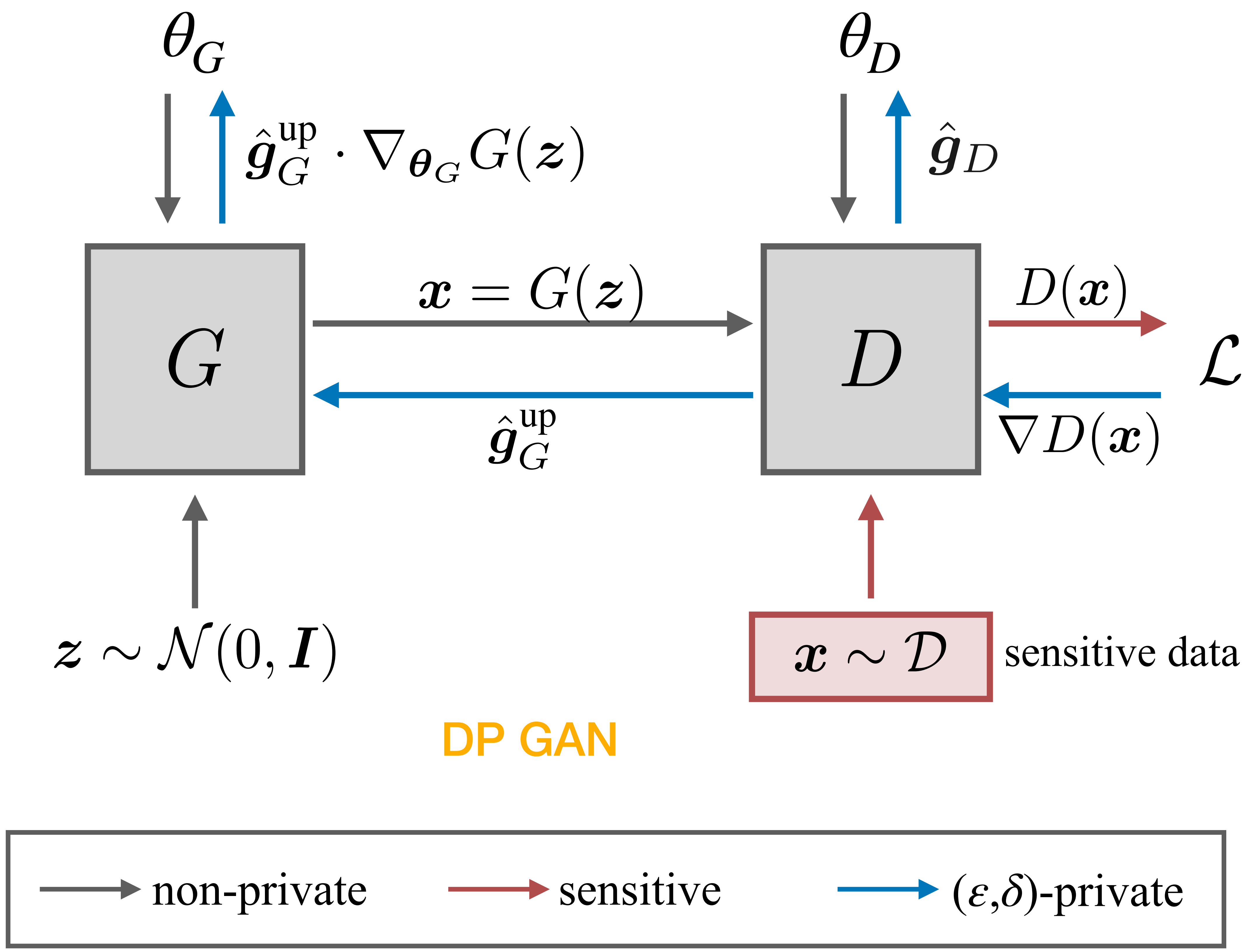
- **Generative adversarial networks (GANs):**
  - Gradient
    $$\boldsymbol{g}^{(t)} := \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}_D, \boldsymbol{\theta}_G)$$
  - Sanitization mechanism
    $$\hat{\boldsymbol{g}}^{(t)} := \mathcal{M}_{\sigma,C}(\boldsymbol{g}^{(t)}) = \mathrm{clip}(\boldsymbol{g}^{(t)}, C) + \mathcal{N}(0, \sigma^2 C^2 \boldsymbol{I})$$
  - Gradient descent step
    $$\boldsymbol{\theta}^{(t+1)} := \boldsymbol{\theta}^{(t)} - \eta \cdot \hat{\boldsymbol{g}}^{(t)}$$




DP GAN

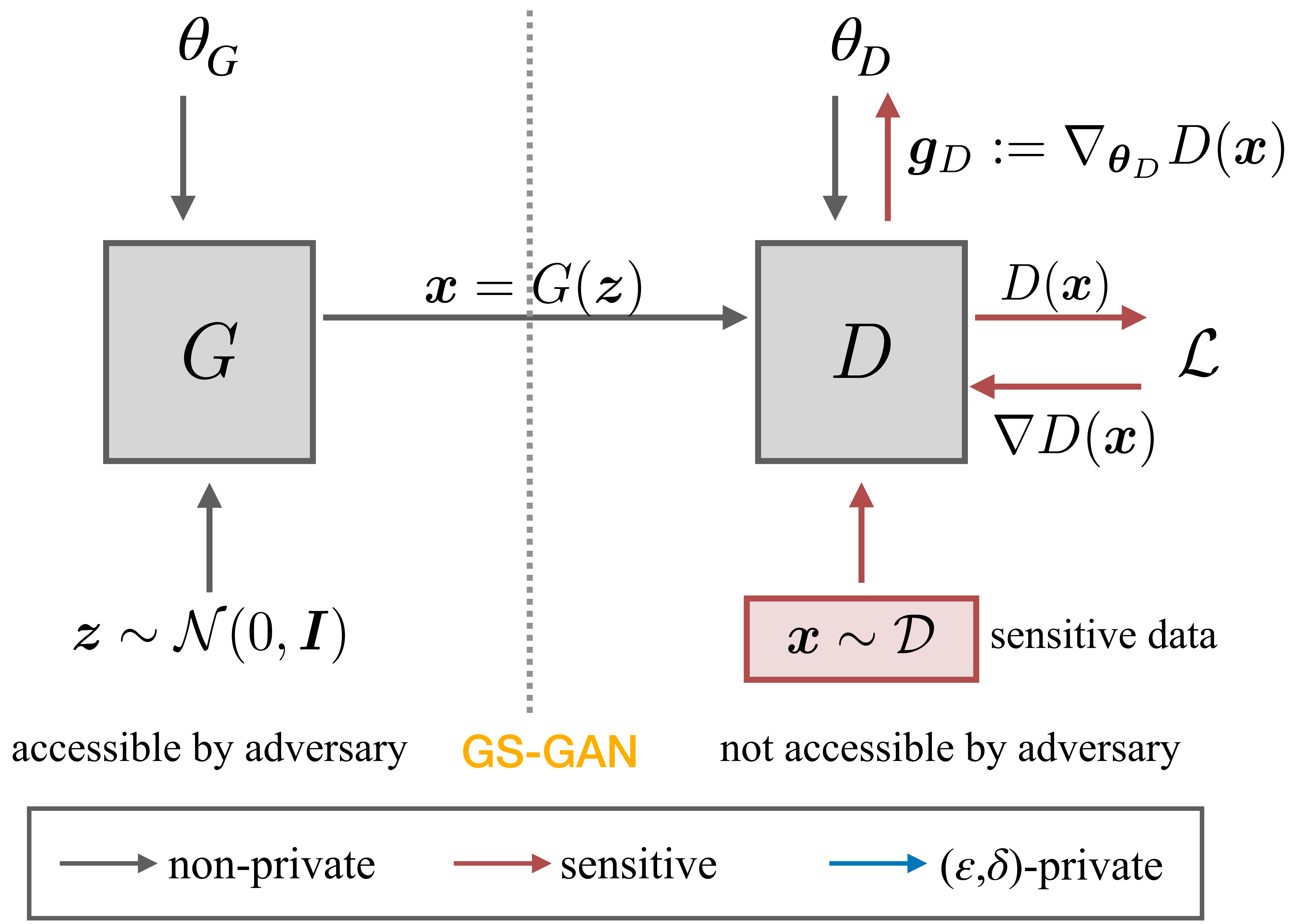
# Existing Solutions

- **Generative adversarial networks (GANs):**
  - Gradient
    $$g^{(t)} := \nabla_{\theta} \mathcal{L}(\theta_D, \theta_G)$$
  - Sanitization mechanism
    $$\hat{g}^{(t)} := \mathcal{M}_{\sigma,C}(g^{(t)})$$
    $$= \text{clip}(g^{(t)}, C) + \mathcal{N}(0, \sigma^2 C^2 I)$$
    clipping bound
  - Gradient descent step
    $$\theta^{(t+1)} := \theta^{(t)} - \eta \cdot \hat{g}^{(t)}$$



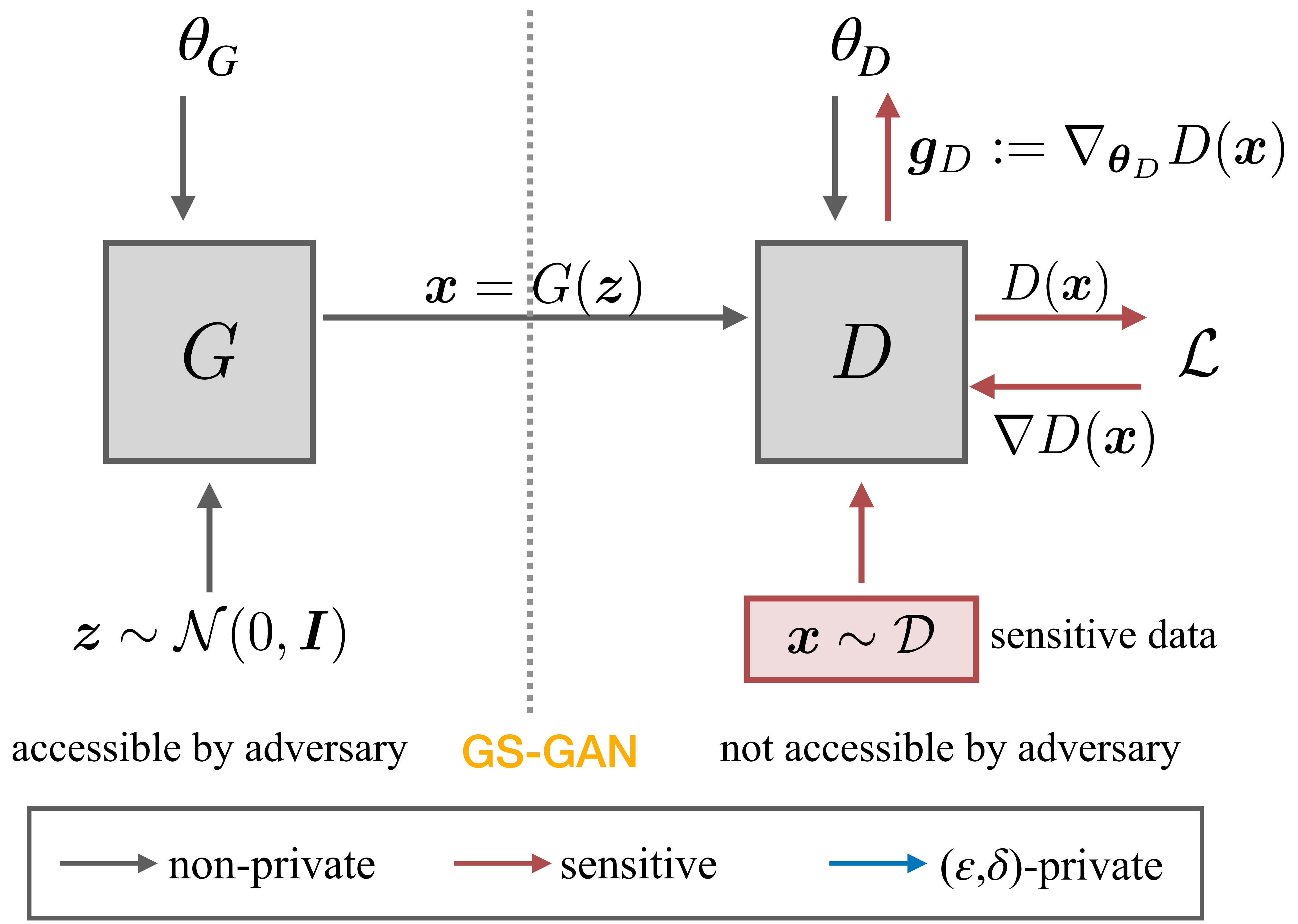
# GS-WGAN: A Gradient-Sanitized Approach for Learning Differentially Private Generators (NeurIPS 2020)

- **Insight:**
  - Only the *generator* need to be publicly-released



# GS-WGAN: A Gradient-Sanitized Approach for Learning Differentially Private Generators (NeurIPS 2020)

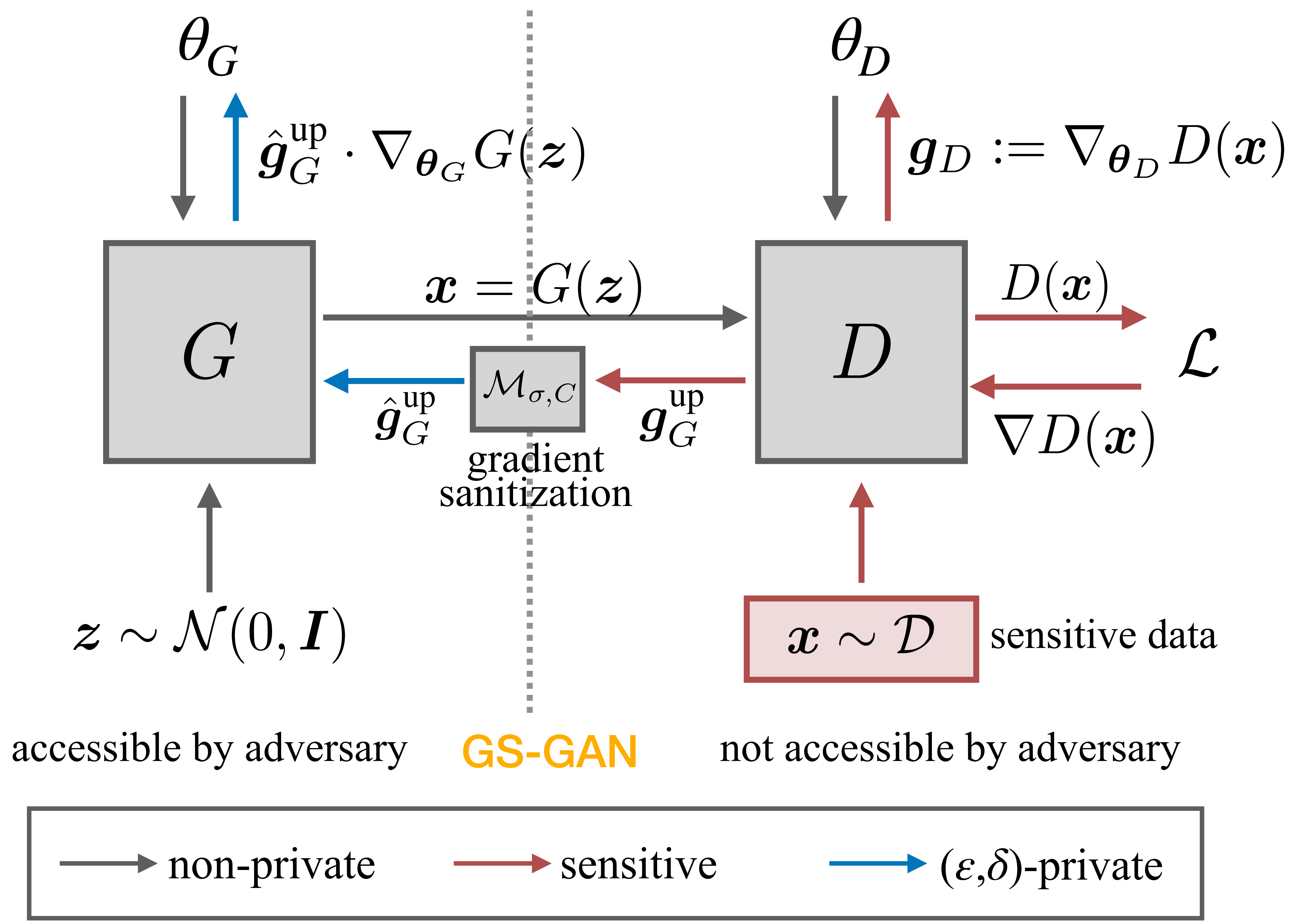
- **Insight:**
  - Only the *generator* need to be publicly-released
- **Our framework:**
  1. Selectively applying sanitization mechanism



# GS-WGAN: A Gradient-Sanitized Approach for Learning Differentially Private Generators (NeurIPS 2020)

- **Insight:**
  - Only the *generator* need to be publicly-released

- **Our framework:**
  1. Selectively applying sanitization mechanism



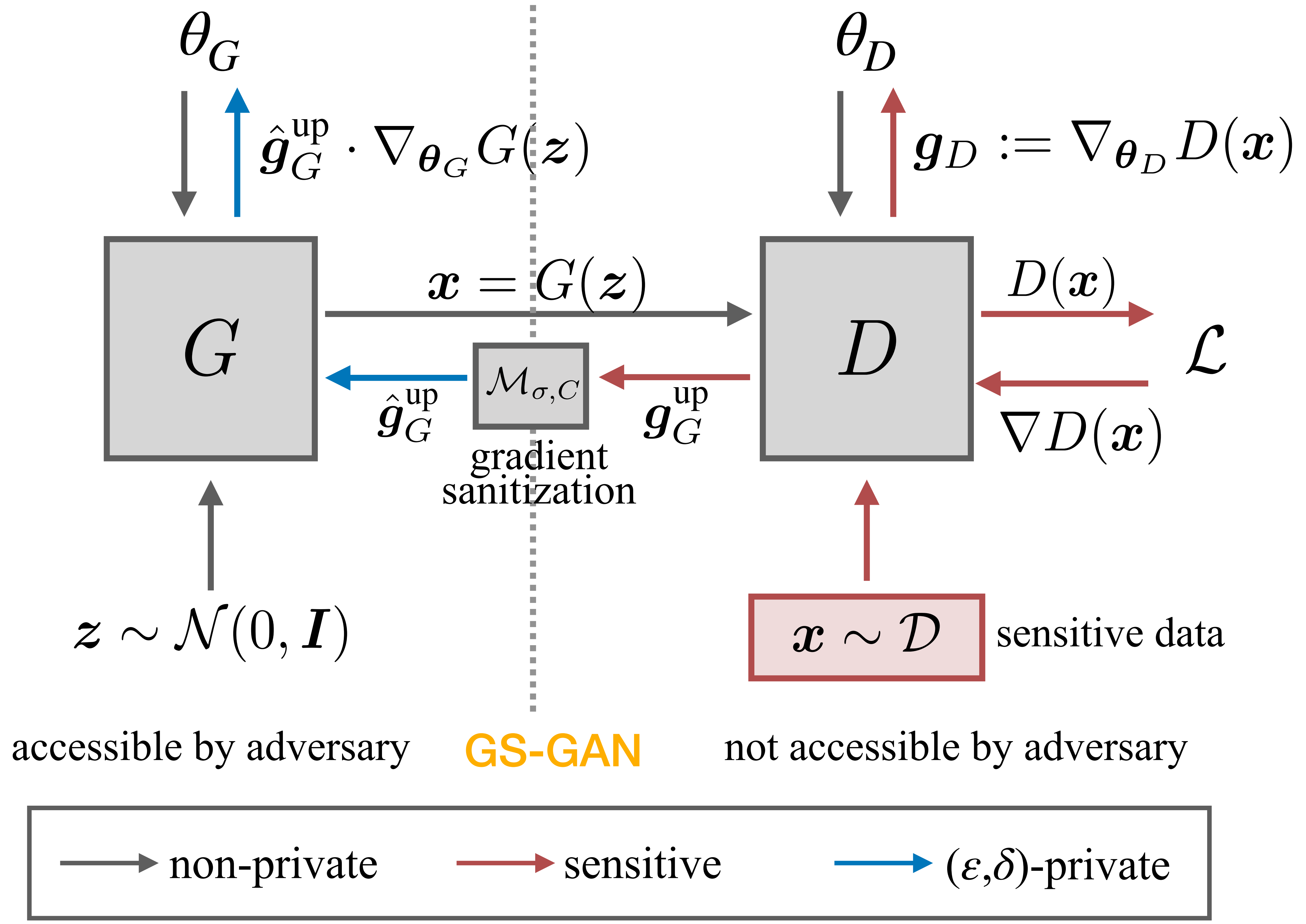
# GS-WGAN: A Gradient-Sanitized Approach for Learning Differentially Private Generators (NeurIPS 2020)

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

- **Insight:**
  - Only the *generator* need to be publicly-released

- **Our framework:**
  1. Selectively applying sanitization mechanism

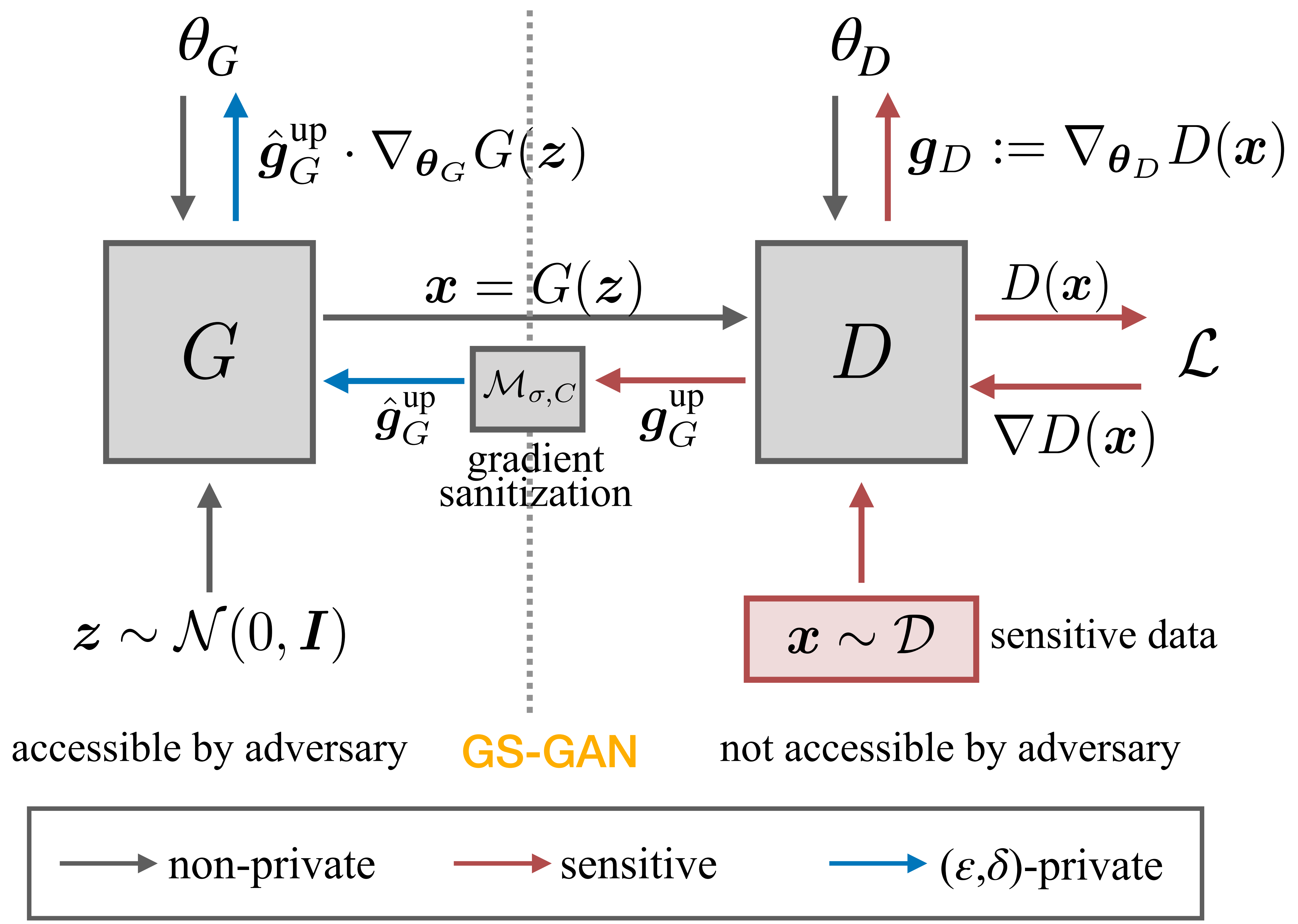
- **Advantages:**
  1. Maximally preserve the true gradient direction



# GS-WGAN: A Gradient-Sanitized Approach for Learning Differentially Private Generators (NeurIPS 2020)

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

- **Insight:**
  - Only the *generator* need to be publicly-released
- **Our framework:**
  1. Selectively applying sanitization mechanism
  2. Bounding sensitivity using Wasserstein distance
- **Advantages:**
  1. Maximally preserve the true gradient direction

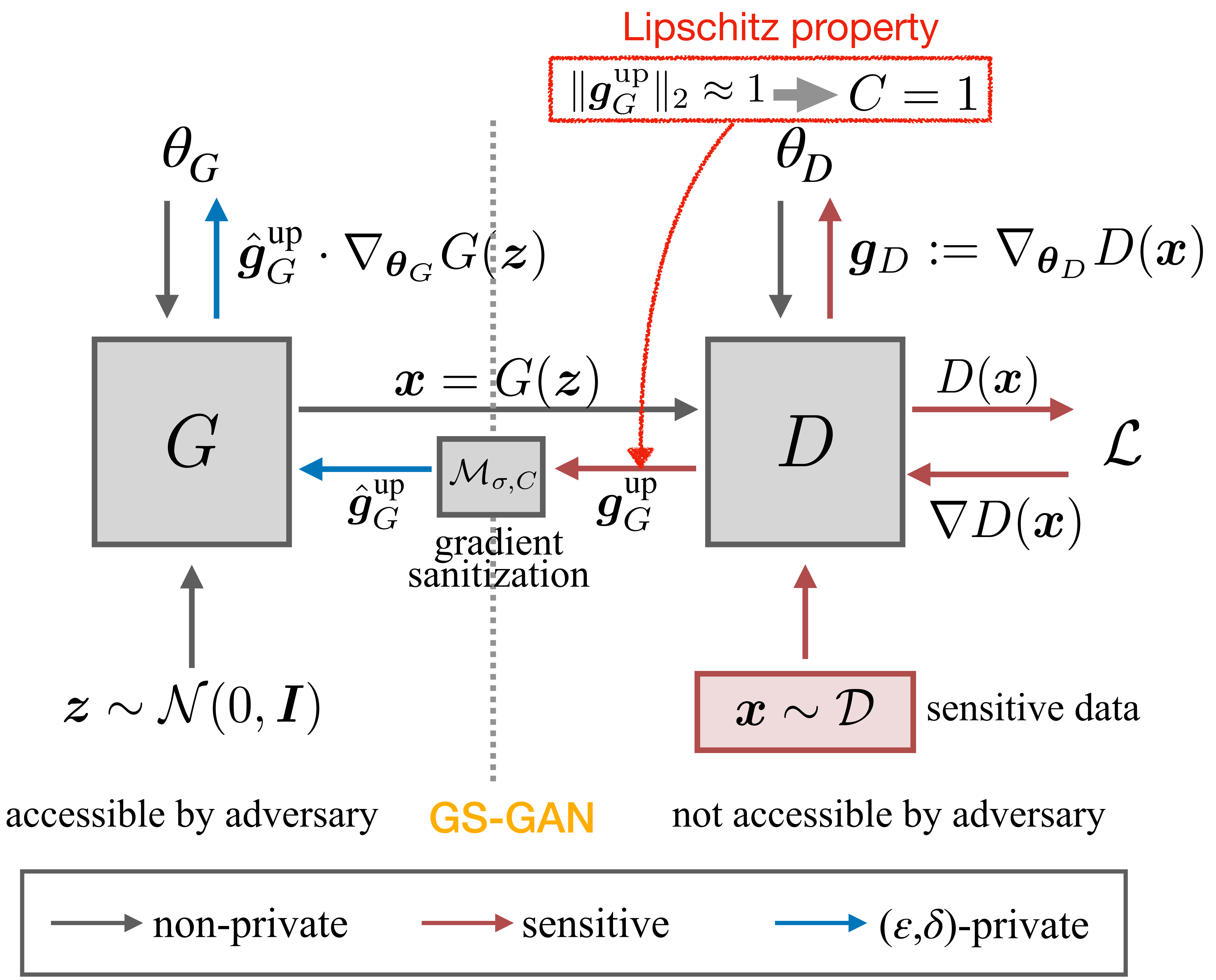


# GS-WGAN: A Gradient-Sanitized Approach for Learning Differentially Private Generators (NeurIPS 2020)

- **Insight:**
  - Only the *generator* need to be publicly-released

- **Our framework:**
  1. Selectively applying sanitization mechanism
  2. Bounding sensitivity using Wasserstein distance

- **Advantages:**
  1. Maximally preserve the true gradient direction

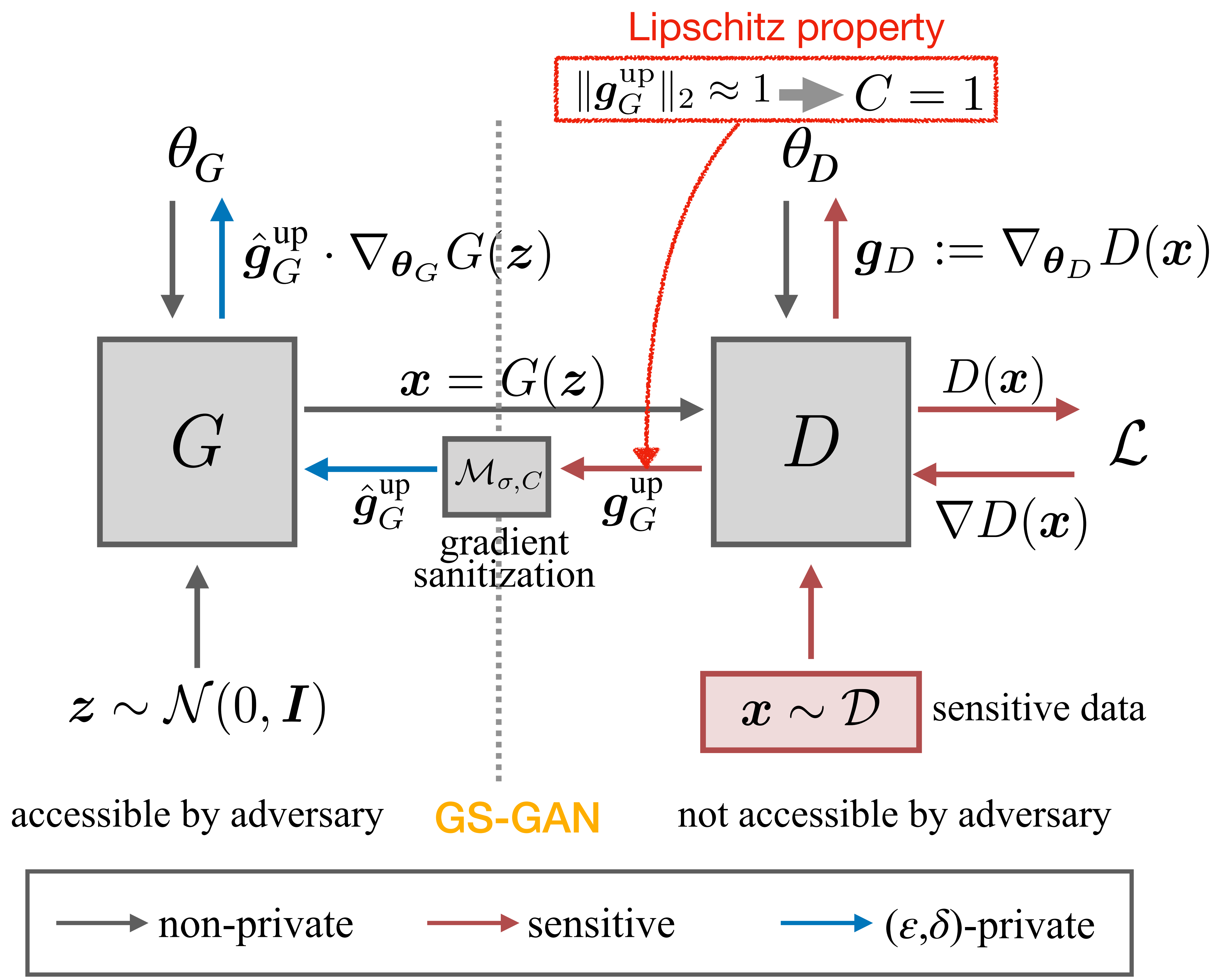


# GS-WGAN: A Gradient-Sanitized Approach for Learning Differentially Private Generators (NeurIPS 2020)

- **Insight:**
  - Only the *generator* need to be publicly-released

- **Our framework:**
  1. Selectively applying sanitization mechanism
  2. Bounding sensitivity using Wasserstein distance

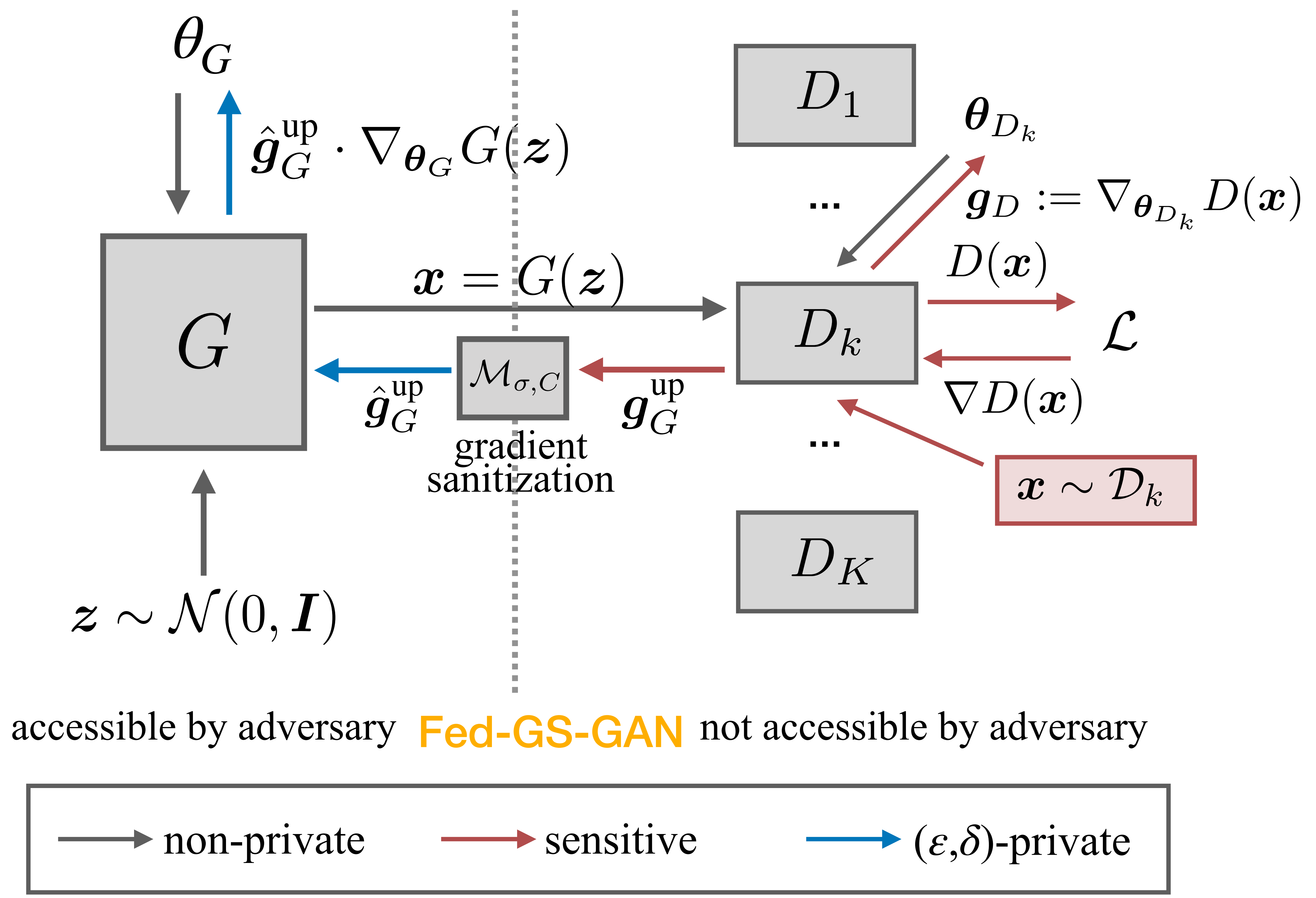
- **Advantages:**
  1. Maximally preserve the true gradient direction
  2. Bypass an intensive and fragile hyper-parameter search for clipping value
  3. Small clipping bias



# GS-WGAN: A Gradient-Sanitized Approach for Learning Differentially Private Generators (NeurIPS 2020)

- **Decentralized (Federated) setting:**
  - Each user train a discriminator on its sensitive dataset locally
  - Communicate the sanitized gradient
- **Advantages:**
  - User-level DP guarantee under an *untrusted* server
  - Communication-efficient (gradients w.r.t. generated samples are *more compact* than gradients w.r.t model parameters[1])

$$\dim(\hat{\boldsymbol{g}}_G^{\text{up}}) \ll \dim(\boldsymbol{\theta}_G) \ll \dim(\boldsymbol{\theta}_G) + \dim(\boldsymbol{\theta}_D)$$




[1] Augenstein et al., "Generative Models for Effective ML on Private, Decentralized Datasets", ICLR 2020

# GS-WGAN: A Gradient-Sanitized Approach for Learning Differentially Private Generators (NeurIPS 2020)

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

- **Adopted and extended by SOTA following works:**
  - Long, Yunhui, et al., "G-PATE: Scalable Differentially Private Data Generator via Private Aggregation of Teacher Discriminators." (*NeurIPS,* 2021)
  - Cao, Tianshi, et al., "Don't Generate Me: Training Differentially Private Generative Models with Sinkhorn Divergence.", (*NeurIPS,* 2021)
  - Wang, Boxin et al., "Datalens: Scalable privacy preserving training via gradient compression and aggregation." (*CCS,* 2021)

Session 7A: Privacy Attacks and Defenses for ML — CCS '21, November 15–19, 2021, Virtual Event, Republic of Korea

**DataLens: Scalable Privacy Preserving Training via Gradient Compression and Aggregation**

**Don't Generate Me: Training Differentially Private Generative Models**

**G-PATE: Scalable Differentially Private Data Generator via Private Aggregation of Teacher Discriminators**

Yunhui Long[1*] Boxin Wang[1*] Zhuolin Yang[1] Bhavya Kailkhura[2] Aston Zhang[1]
Carl A. Gunter[1] Bo Li[1]
[1] University of Illinois, Urbana Champaign [2] Lawrence Livermore National Laboratory
{ylong4, boxinw2, zhuolin5, lzhang74, cgunter, lbo}@illinois.edu

**Abstract**

Recent advances in machine learning have largely benefited from the massive accessible training data. However, large-scale data sharing has raised great privacy concerns. In this work, we propose a novel privacy-preserving data Generative model based on the PATE framework (G-PATE), aiming to train a scalable differentially private data generator which preserves high generated data utility. Our approach leverages generative adversarial nets to generate data, combined with private aggregation among different discriminators to ensure strong privacy guarantees. Compared to existing approaches, G-PATE significantly improves the use of privacy budgets. In particular, we train a student data generator with an ensemble of teacher discriminators and propose a novel private gradient aggregation mechanism to ensure differential privacy on all information that flows from teacher discriminators to the student generator. In addition, with random projection and gradient discretization, the proposed gradient aggregation mechanism is able to effectively deal with high-dimensional gradient vectors. Theoretically, we prove that G-PATE ensures differential privacy for the data generator. Empirically, we demonstrate the superiority of G-PATE over prior work through extensive experiments. We show that G-PATE is the first work being able to generate high-dimensional image data with high data utility under limited privacy budgets ($\varepsilon \leq 1$). Our code is available at https://github.com/AI-secure/G-PATE

**1 Introduction**

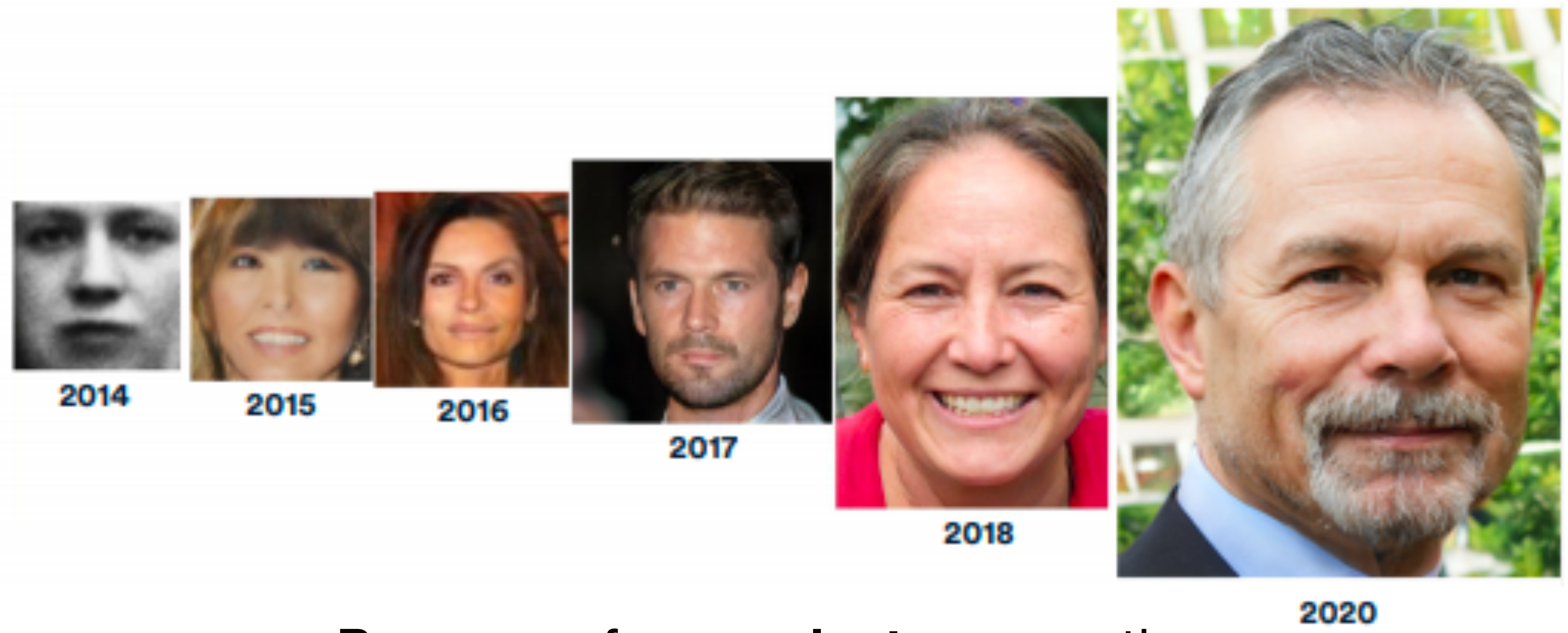
Machine learning has been applied to a wide range of applications such as face recognition [30, 39, 21, 22], autonomous driving [26], and medical diagnoses [8, 20]. However, most learning methods rely on the availability of large-scale training datasets containing sensitive information such as personal photos or medical records. Therefore, such sensitive datasets are often hard to be shared due to privacy concerns [40]. To handle this challenge, data providers sometimes release synthetic datasets produced by generative models learned on the original data. Though recent studies show that generative models such as generative adversarial networks (GAN) [14] can generate synthetic records that are indistinguishable from the original data distribution, there is no theoretical guarantee on the privacy protection. While privacy definitions such as differential privacy [9] and Rényi differential privacy [27] provide rigorous privacy guarantee, applying them to synthetic data generation is nontrivial.

Recently, two approaches have been proposed to combine differential privacy with synthetic data generation: DP-GAN [35] and PATE-GAN [37]. DP-GAN modifies GAN by training the discriminator using differentially private stochastic gradient descent. Though it achieves privacy guarantee due to

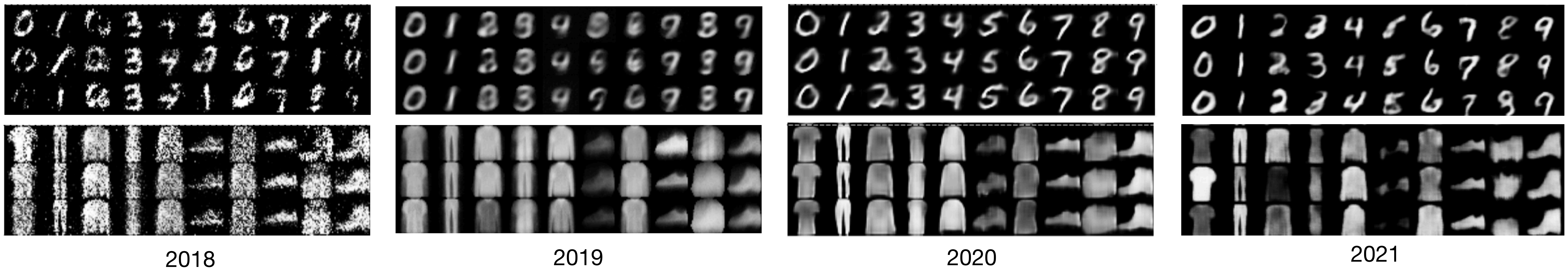
*Equal contribution.

35th Conference on Neural Information Processing Systems (NeurIPS 2021)

# Challenges




Progress of **non-private** generation




Progress of **private** generation $(\varepsilon, \delta) = (10, 10^{-5})$

**Saturated? Problem too hard?**

# Challenges

- Fitting the complete high-dimensional data distribution is complicated
    - Deep generative models are **data demanding**
    - Privacy constraints
- No enough data to solve such a difficult problem ☹

# Private Set Generation with Discriminative Information (NeurIPS 2022)

- **Existing approaches:**
  - Aim at fitting the complete data distribution
  - Optimize deep generative models
  - Suboptimal utility: <85% for MNIST with $(\varepsilon,\delta)=(10, 10^{-5})$

- **Our approach:**
  - Target at common downstream tasks (e.g., classification)
  - Directly optimize a set of representative samples
  - ~10% downstream test accuracy improvement over SOTA

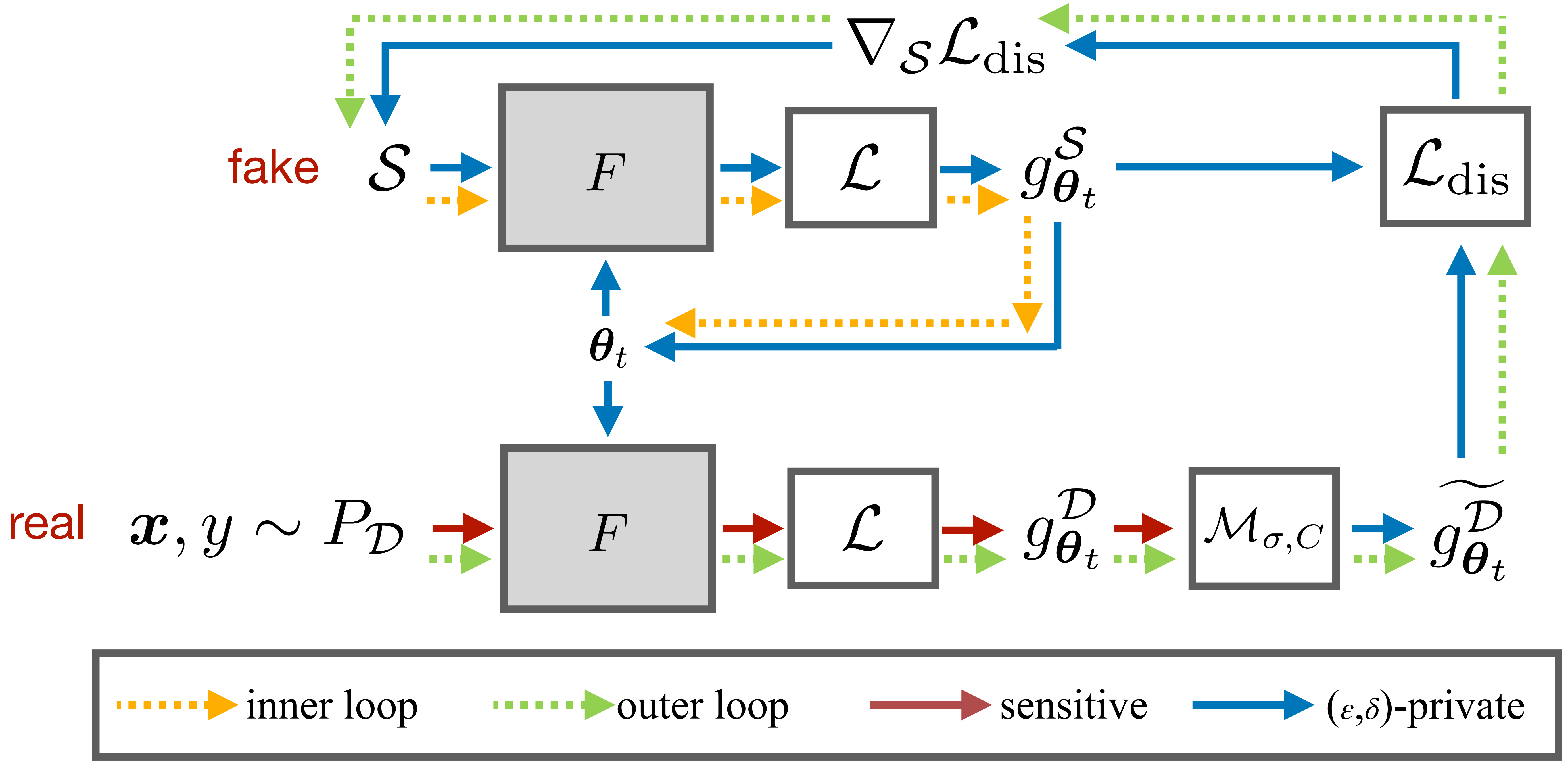
**Generally easier**

**Better convergence**

**Useful samples**

# Private Set Generation with Discriminative Information (NeurIPS 2022)

- **Target:**
  - Optimize for training downstream Neural Network classifier
- **Basic idea:**
  - Gradient-based **coreset generation**[1,2]
  - DP stochastic gradient descent (DP-SGD)

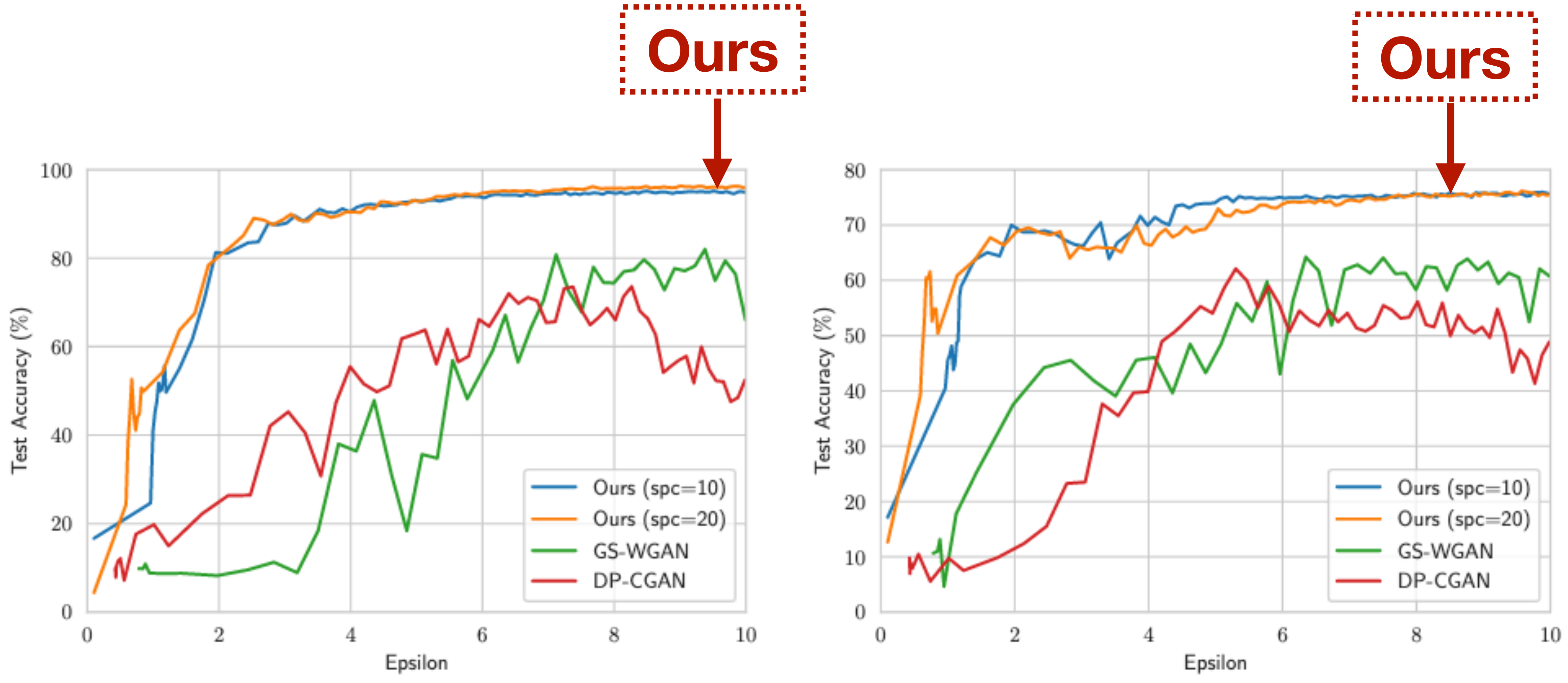



[1] Zhao, Bo, et al., "Dataset condensation with gradient matching.", *ICLR,* 2021.
[2] Zhao, Bo, et al., "Dataset condensation with differentiable siamese augmentation.", *ICML*, 2021

# Private Set Generation with Discriminative Information (NeurIPS 2022)

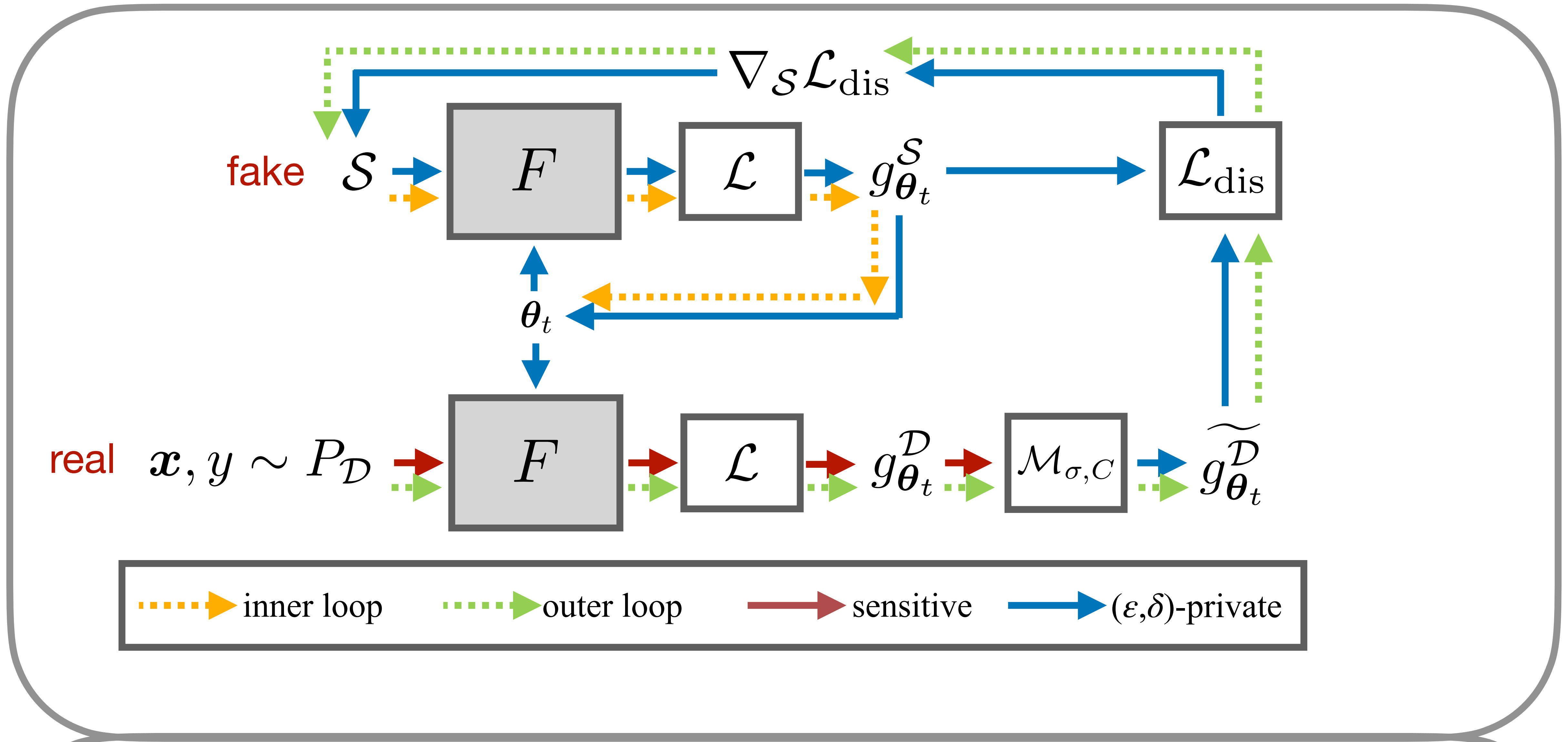
- **Comparison to SOTA:**
  - Utility for downstream classification task (train on synthetic; test on real)
  - Convergence rate

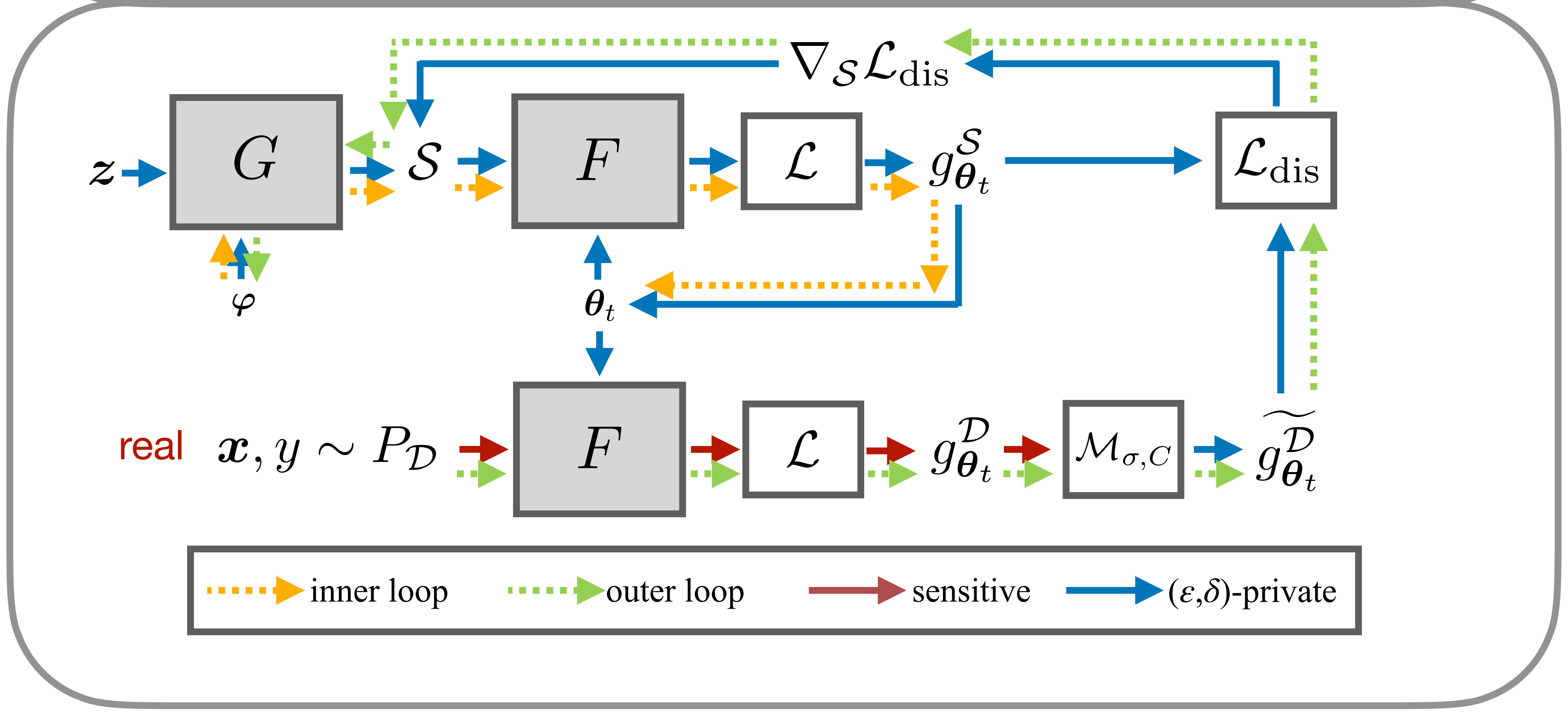
| | MNIST | | | | | | FashionMNIST | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ConvNet | LeNet | AlexNet | VGG11 | ResNet18 | MLP | ConvNet | LeNet | AlexNet | VGG11 | ResNet18 | MLP |
| Real | 99.6 | 99.2 | 99.5 | 99.6 | 99.7 | 98.3 | 93.5 | 88.9 | 91.5 | 93.8 | 94.5 | 86.9 |
| DP-CGAN | 50.2 | 52.6 | 52.1 | 54.7 | 51.8 | 54.3 | 50.2 | 52.6 | 52.1 | 54.7 | 51.8 | 54.3 |
| GS-WGAN | 84.9 | 83.2 | 80.5 | 87.9 | 89.3 | 74.7 | 54.7 | 62.7 | 55.1 | 57.3 | 58.9 | 65.4 |
| DP-Merf | 85.7 | 87.2 | 84.4 | 81.7 | 81.3 | 85.0 | 72.4 | 67.9 | 64.9 | 70.1 | 66.7 | **73.1** |
| Ours (spc=10) | 94.9 | 91.3 | 90.3 | 93.6 | **94.3** | 86.1 | 75.6 | **68.0** | **66.2** | 74.7 | **72.1** | 62.8 |
| Ours (spc=20) | **95.6** | **93.0** | **92.3** | **94.5** | 94.1 | **87.1** | **77.7** | **68.0** | 59.1 | **76.8** | 70.8 | 62.2 |



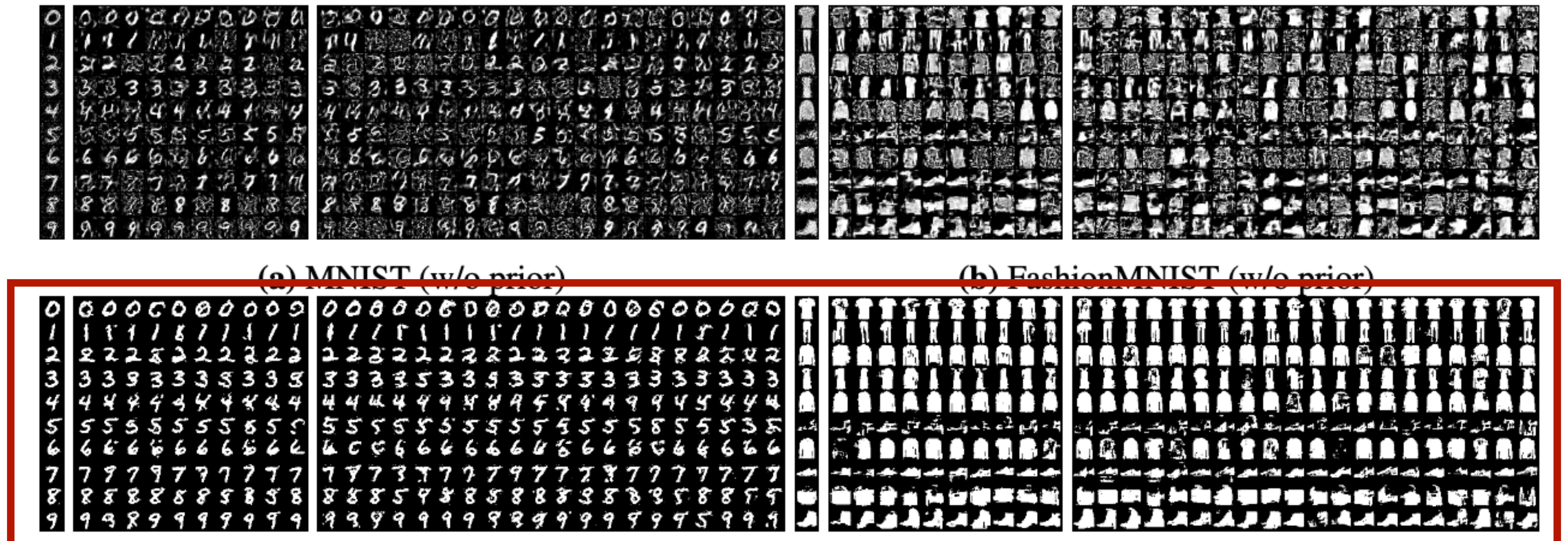

(a) MNIST (b) FashionMNIST

# Private Set Generation with Discriminative Information (NeurIPS 2022)




- Deep generator structure result in:
  - Better visual quality
  - Slow convergence
  - Sub-optimal downstream utility



(a) MNIST (w/o prior) (b) FashionMNIST (w/o prior)

(c) MNIST (with prior) (d) FashionMNIST (with prior)

**with generative model**

| | MNIST | | | FashionMNIST | | |
|---|---|---|---|---|---|---|
| | 1 | 10 | 20 | 1 | 10 | 20 |
| w/o prior | 81.4 | **94.9** | **95.6** | **66.7** | **75.6** | **77.7** |
| with prior | **88.2** | 92.2 | 90.6 | 63.0 | 70.2 | 70.7 |

# In summary:

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

- **Privacy-preserving Generation is important**
  - Flexibility & Transparency: downstream analysis, reproducible research
  - Applications: federated learning
- **Privacy-preserving Generation is non-trivial:**
  - Exploit the progress in general generative modeling
  - Co-design of private- and non-private models
  - Make better usage of “prior knowledge”
    - Task (downstream model)
    - Data distribution

# Thanks for your attention!

**Presenter:** Dingfan Chen
**Supervisor:** Prof. Dr. Mario Fritz
**Affiliation:** CISPA – Helmholtz Center for Information Security